Enhancing Media Personalization by Extracting Similarity Knowledge from Metadata

Butkus, Andrius

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Enhancing Media Personalization by Extracting Similarity Knowledge from Metadata

Andrius Butkus

Kongens Lyngby 2008
IMM-PHD-2008-198

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

mano seneliui Juozui Knispeliui

Summary

The world of media today can be characterized by us being exposed to vast amounts of content, both produced professionally and user generated. Ever since the digital technologies in the form of computers and video cameras have diminished the production costs and the Internet has significantly lowered the costs of distribution, we became more and more overwhelmed with the choice of media. In such conditions the focus falls on the available mechanisms to filter and recommend media to users, thus resulting in the growing need for personalization.

Media personalization is a complex process with many interrelated parts – recommendation engines, content metadata, contextual information and user profiles. In the center of any type of recommendation lies the notion of *similarity*. The most popular way to approach similarity is to look for the feature overlaps. This results often in recommending only "more of the same" type of content which does not necessarily lead to the meaningful personalization. Another way to approach similarity is to find a similar underlying meaning in the content.

Aspects of meaning in media can be represented using Gärdenfors *Conceptual Spaces* theory, which can be seen as a cognitive foundation for modeling concepts. Conceptual Spaces is applied in this thesis to analyze media in terms of its dimensions and knowledge domains, which in return defines properties and concepts. One of the most important domains in terms of describing media is the emotional one, especially when we talk about the contents of music. Therefore the main focus in the thesis is how to extract such emotional information from media, and how to use it to enhance media personalization.

This dissertation proposes a novel method to extract emotional information from text (unstructured metadata) using *Latent Semantic Analysis* (one of the unsupervised machine learning techniques). It presents three separate cases to illustrate the similarity knowledge extraction from the metadata, where the emotional components in each case represents different abstraction levels – *genres*, *synopsis* and *lyrics*. The emotional value is extracted by first creating a conceptual space for emotions based on a semantic differential which divides the underlying plane along two psychological dimensions – *arousal* and *valence*. Then the space is divided into regions serving as emotional markers – a selection of affective terms. After that LSA is used to calculate the cosine similarity between the text (synopsis or lyrics) and each of the chosen affective terms. As a result we can plot emotional correlation in the content as patterns, which we can then use to find emotional similarity among media items.

By being able to compare media items on the basis of their emotional patterns, we add a new level to how we can evaluate the similarity between two media items. Which in return might improve media recommendation since it provides a novel approach to recommendation that goes beyond traditional genre boundaries, and thereby improves media personalization.

Resumé

Den verden af medier vi i stigende grad er blevet en del af er karakteriseret ved at vi eksponeres til store mængder af indhold, dels produceret professionelt såvel som skabt af brugerne selv. Lige siden digital teknologi i form af computere og video kameraer formindskede produktionsomkostningerne og internettet sænkede udgifterne til distribution, er vi i stigende grad blevet overvældet af valgmulighederne indenfor medieindhold. I den forbindelse har interessen samlet sig om de mekanismer der gør det muligt at filtrere og anbefale indhold til brugerne, der afspejler et voksende behov for personalisering.

Personalisering af medieindhold er en kompleks process med mange komponenter der er afhængige af hinanden – den bagvedliggende logik til at anbefale indhold, det metadata der beskriver indholdet, kontekstuel information samt bruger profiler. I alle typer af anbefaling af indhold er det centrale begreb *similaritet*. Den oftest benyttede måde til at definere similaritet er at finde fælles karakteristika. Det resulterer ofte kun i anbefaling af "mere af samme slags" indhold, hvilket ikke nødvendigvis resulterer i meningsfuld personalisering. En anden tilgang til similaritet kunne være istedet at finde en fælles dybere mening der ligger til grund for indholdet.

Aspekter af mening i medieindhold kan repræsenteres med brug af Gärdenfors *Conceptual Spaces* teori, der kan opfattes som et cognitivt fundament for modellering af koncepter. Conceptual Spaces anvendes i denne afhandling til at analysere indhold i forhold til dimensionalitet og domæner, hvilket igen ligger til grund for definition af karakteristika og koncept. Et af de vigtigste domæner i forhold til at beskrive medieindhold er det emotionelle, ikke mindst når vi taler om indholdet i musik. Fokus i denne afhandling er derfor hvordan vi kan ekstrahere denne emotionelle information fra multimedier, og derefter anvende

den til forbedret personalisering af indholdet.

Denne afhandling præsenterer en ny måde til at ekstrahere emotionel information fra tekst (ustruktureret metadata) med anvendelse af LSA *Latent Semantisk Analyse* som usuperviseret machine learning teknik. Tre separate case's bruges til at illustrere hvordan elementer af similaritet kan ekstraheres fra metadata, hvor de emotionelle komponenter repræsenterer forskellige niveauer af abstraktion – *genrer*, *synopsis* og *sangtekster*. De emotionelle komponenter modelleres ved at definere et konceptuelt emotionelt rum udfra et semantisk differentiale, der opdeler det underliggende plan i to psykologiske dimensionser – *arousal* og *valence*. Rummet opdeles i regioner der fungerer som emotionelle sensorer – et udvalg af affektive termer. Derefter anvendes LSA til at udregne den cosinus similaritet der er imellem tekst (synopsis eller sangtekster) og hver af de udvalgte affektive termer. Som resultat kan vi udtrykke emotionel correlation i indholdet som mønstre, som derefter kan bruges som grundlag for at finde andet medieindhold der reflekterer de samme underliggende følelser.

Ved at gøre det muligt at sammenligne indhold udfra emotionelle mønstre, tilføjer vi et nyt plan til at bestemme similaritet i medieindhold. Hvilket igen vil ændre hvad der anbefalies, da det giver en ny tilgang til at præsentere brugeren for indhold der ikke er begrænset af traditionelle genre grænser, og demed forbedre personaliseringen af medieindhold.

Preface

This thesis was prepared at Informatics Mathematical Modelling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

In todays media world we really do not lack variety. What we do lack is the ability to efficiently find media we like. And if we find what we like, we have no idea of what else might be out there that we may like even more. This thesis confronts the question of how to increase similarity knowledge in order to improve recommendation quality, and outlines a novel approach to automatically extract latent semantics from descriptions of broadcast TV programs and song lyrics as a basis for modeling the emotional context of media

The thesis consists of a summary report and a collection of the five research papers written during the period 2005–2008, and elsewhere published.

Lyngby, April 2008

Andrius Butkus

Publications

Papers included in the thesis

[A] Andrius Butkus and Michael Kai Petersen. Semantic Modeling Using TV-Anytime Genre Metadata. *Book Chapter in "Interactive TV: A shared experience, Proceedings of EuroITV 2007 (ISBN: 978-3-540-72558-9)"*, Springer-Verlag, Berlin 2007.

[B] Michael Kai Petersen and Andrius Butkus. Modeling Moods in BBC Programs Based on Emotional Context. *Book Chapter in "Lecture Notes in Computer Science", Proceedings of EuroITV 2008*, Springer-Verlag, Berlin 2007

[C] Michael Kai Petersen and Andrius Butkus. Semantic Contours in Tracks Based on Emotional Tags. *In Proceedings of "Computer Music Modeling and Retrieval"*, 2008.

[D] Michael Kai Petersen and Andrius Butkus. Extracting Moods from Songs and BBC Programs Based on Emotional Context. *In Journal "International Journal of Digital Multimedia Broadcasting"*, 2008.

[E] Michael Kai Petersen, Lars Kai Hansen, Andrius Butkus and Martin Schwartz. Emotional Vectors: Modeling Media from Cognitive Components. *Submitted to the "Journal of Multimedia Systems"*, 2008.

Other journal papers or conference contributions published during the preparation of the thesis

- Michael Kai Petersen and Andrius Butkus. Modeling Emotional Context From Latent Semantics. *In Proceedings of "The 1st International Conference on Designing Interactive User Experiences for TV and Video"*, 2008.
- Michael Kai Petersen and Andrius Butkus. Extracting Patterns from Tracks Using Emotional Tags. *In Proceedings of "The 2nd International Workshop on Learning Semantics of Audio Signals"*, 2008.
- Andrius Butkus. Media Personalization Using TV-Anytime Phase 2 *In Proceedings of "The 3rd International CICT Conference, Mobile and Wireless Content, Services and Networks"*, 2006.
- Andrius Butkus and Henning Olesen Value Creation and New Business Opportunities by Means of Personalization in Future Converged Services. *In Proceedings of "IST Mobile and Wireless Communications Summit"*, 2006.
- Andrius Butkus. New Business opportunities for DVB-H Personalized Services Using CRID *In Proceedings of "The 2nd International CICT Conference, Next Generation Broadband, Content and User Perspectives"*, 2005.

Acknowledgements

I would like to take this opportunity to first and foremost thank my colleague, friend and soulmate Michael Kai Petersen. We have had a very close collaboration, the results of which can be easily seen form the publication list. We challenged each other on a daily basis while having a lot of fun in the process. I can not express enough how valuable our daily discussions were. But most of all I would like to thank him just for being there when I needed him most.

I have to thank my girlfriend Emilie ♡ for being unbelievably patient with me during all this complicated and messy time. Her selfless love and daily support has helped me to make it through without going completely crazy. I will remain eternally grateful for every single little thing she did to make sure I was OK, whether it was a cup of coffee late at night or a big hug when I was down... Thank you Emi...

And where would I be without the most cheerful, energetic and optimistic member of my wonderful support crew – Wivi. She always found ways to inspire me and also bring out the best in me. I can already see her sweet smiling eyes saying "I told you that you can do it..." I did. But not without your help, Wivi.

Last but definitely not least I would like to say a big thank you to my family who believed in me all the time no matter what. It felt like they were with me every step along the way, even though physically they all were in another country. I want to say a special thank you to my granddad to whom I dedicate this book. During these last three years our special connection got even more special crossing all conceivable boundaries. His spiritual involvement and excitement about my work made it so much sweeter to finally finish it.

Contents

Chapter 1

Introduction

This thesis analyzes media personalization problem. Media personalization boils down to automatically filtering media based on the given parameters (user preferences and context). Filtering takes either a form of search or recommendations. In most cases those two are combined, and that means that every time we search for something we also get recommendations which if done properly are highly effective tool for the users to explore the ever expanding media world. Recommendations are usually based on finding similarities among media. Therefore the thesis focuses on the media descriptions, with the main goal to improve media retrieval by introducing new ways to increase knowledge about content and use it to enable more meaningful recommendations.

The thesis builds on the cognitive approaches to categorization to approach the "similarity" problem. Conceptual Spaces [Gärdenfors, 2000] is used as the framework to represent knowledge and to model a conceptual space for media. The main challenge there is to get down to the fundamental dimensions that represent how we perceive media. Since knowledge extraction from the raw media (songs, movies, books, etc) is not advanced enough to meet our needs, the only material that we can rely on are the media descriptions (from categories and keywords all the way to the synopsis and user reviews, and even song lyrics), and since there are different ways to describe media, different methodologies need to be used to process it. On the methodology side the thesis builds on a number of methods and and techniques. The main techniques are Latent Semantic Analysis

[Deerwester et al., 1990] used to extract latent semantic knowledge from the text and semantic differential [Osgood et al., 1957] for presenting dimensions for building emotional space. This introductory chapter raises a need for such analysis and presents the problem of media personalization. It also introduces the main elements of theory and methodology.

1.1 Background and focus

There has never been so many new opportunities in the media world as it it today. We are in a transition period coming from the world dominated by mass media, professional broadcast and blockbuster hits and entering the world of the Internet, unlimited choice and user generated content. It is not a smooth evolution but rather a real paradigm shift. And as any other transition period it presents a number of challenges and holds many new opportunities for those who are flexible and realize the necessity to adapt. The key, much like in any business, is to know the rules of the game. This section explores the media world starting by showing how it used to be and why it was like that. After that the key factors which changed our economy are presented with examples of how they affected media industry in particular. Finally the new rules are presented that drive the media world today.

1.1.1 Hit-driven economy

Historically, many markets have traditionally been dominated by a few best-selling and featured products [Brynjolfsson et al., 2007]. This applies to all kinds of markets, from physical goods like clothes, cars or food products, to information goods like movies and music. It was not because the consumers wanted it that way, but because of the scarce resources on the supply side of the chain, or as Chris Anderson (editor-in-chief of Wired magazine) puts it "any of our assumptions about popular taste are actually artifacts of poor supply-and-demand matching - a market response to inefficient distribution". As a discipline, economics primarily deals with efficient distribution of scarce resources. Scarcity has been the key concept in economics for years. The two of the main scarcity functions of traditional economics were the marginal costs of manufacturing and the distribution. On the manufacturing side if you have limited resources, it is only natural to produce things that will generate the biggest revenues. Similarly if you have limited distribution resources, it is the most efficient approach to distribute only the goods that sell best. And no matter what industry we look at, they all had limited distribution resources. Industries

that are based on the physical goods will always have limited resources, this is just the reality of the physical world. But the information goods are fundamentally different from the physical ones, nevertheless media used to suffer from the constrains of the physical world as well, mostly due to it's resource-scarce distribution methods.

Those two bottlenecks (production and distribution) were the main reasons limiting consumer's access to more variety. Most of the industries, especially the media industry, were focused on the questions *What is the next big thing?*, *What will sell best?*, and therefore *What is worth producing and distributing?*. Markets were also very limited on how much audience they could reach, which rephrases the question into *What will sell best in this particular area?*. The area of reach highly depended on the market, but in most cases it was very limited. So as a bookstore manager you would inevitably think what books will sell best in your few kilometer radius neighborhood. Hit-driven economics have created an age without enough room to carry everything for everybody. "Not enough shelf space for all the CDs, DVDs, and games produced. Not enough screens to show all the available movies. Not enough channels to broadcast all the TV programs, not enough radio waves to play all the music created, and not enough hours in the day to squeeze everything out through either of those sets of slots". [Anderson, 2004]

Since it's invention the broadcast was the dominant distribution platform for all sorts of media - TV, radio, press, etc. The great thing about broadcast is that it can bring one show to millions of people with unmatched efficiency. But it can't do the opposite - bring millions of shows to one person each. The economics of broadcast required big hits to catch huge audiences. The radio spectrum can carry only so many stations, and a coaxial cable only so many TV channels. And there are only 24 hours of programming per day.

If there is one word to describe the media world of the last 100 years, it would be *hits*. Movies, music, books - all those markets were driven by blockbuster hits that appealed to the masses and sold by millions. The media world was much simpler, smaller, easier to manage but the most importantly it was far more limited. It was the golden age of mainstream. Most people watched the same TV shows, listened to the same music broadcasted on one of a few local radio stations. No wonder that the average top ten TV show viewing in 1950 averaged 44.8% audience, compared with only 13.4% in 2005 [Wang et al., 2006] and that none of the top 5 all time best selling music albums were produced after 1983.

We had no idea of what we are be missing, and even if we knew that media was not likely to be within reach. Even today the scarcity of the distribution is one of the most limiting factor limiting us for greater variety. When we go

to a Blockbuster to rent a movie or to a huge record store for some new music, we might think that we have lots of selection but are seeing only the tip of the iceberg. More than 99% of music albums on the market today are not available in Wallmart (the biggest physical distributor of music), from over 3,7 million published books the average Barnes&Noble carries only 130,000 titles, of more that 200,000 films, TV shows, documentaries and other video that have been released commercially the average Blockbuster carries just 3,000. The same goes for independent films. In 2004, nearly 6,000 movies were submitted to the Sundance Film Festival. Of those, 255 were accepted, and just two dozen have been picked up for distribution; to see the others, you had to be there. The same is for any other leading retailer - from books to kitchen fittings. The vast majority of products are not available at a store near you.[Anderson, 2004].

TV produces more content than any other media industry. There are an estimated 31 million hours of original TV content produced worldwide each year [Anderson, 2006]. Radio produces more hours, but since most radio is music, thus available elsewhere, it does not reach the levels of TV in terms of original content. But only a tiny fraction of all this content is available to the user. The main reason for that is the scarcity of distribution to make everything available live, and the legal issues standing in the way of making everything available on demand.

The economical reality is that where the opportunity cost of inventory storage and distribution is high, only the most popular products are sold. In a nutshell, that was the way we were looking at markets for the last century. “Every retailer has its own economic threshold, but they all cut off what they cary somewhere. Things that are likely to sell in the necessary numbers gets carried, things that aren’t, don’t.” [Anderson, 2006]

1.1.2 The New Economy?

In the early 1990s several things happened that gave birth to the new approach to economy (it was later called - information economy) and fundamentally changed the way business is done, especially in the media world. There are three main forces that affect all markets: technology, economics and regulations. In this case it was the technology that was the driving force and these are the main technological factors:

- Invention of the global information network - the Internet - made it possible to access information almost anywhere and significantly reduced the information delivery cost.

- Development of the new data transmission technologies (DSL, WLAN, WiFi, WiMAX, etc) resulted in the increase of the available bandwidth and allowed rich media to be delivered via internet
- Convergence of previously different platforms sharing IP (Internet Protocol) as a core delivery protocol (convergence of TV, radio, mobile, games, etc).
- Decrease of prices of all ICT equipment resulted in an ever-growing number of devices, with increasing amounts of memory and processing capacity. And in return cheap devices allowed individuals inexpensively create loads of content (YouTube, Flickr, etc).
- Media shift from analog to digital opened up possibilities for indexing and searching.
- Return channel in internet and mobile platforms allows to deliver interactive media and enables personalized services.

Venture capitalist David Hornik says: "The basic idea is that incredible advances in technology have driven the cost of things like transistors, storage, bandwidth, to zero. And when the elements that make up a business are sufficiently abundant as to approached free, companies appropriately should view their businesses differently than when resources were scarce (the Economy of Scarcity). They should use those resources with abandon, without concern for waste. That is the overriding attitude of the Economy of Abundance - don't do one thing, do it all; don't sell one piece of content, sell it all; don't store one piece of data, store it all." [Hornik, 2006].

In other words it was the advances in electronic industry that enabled cheap devices and thus democratized the tools of production, the Internet on the other hand has democratized the tools of distribution.

The new situation is so fundamentally different from what we had before that it has inspired many economists to rethink their views on economy and resulted in numerous influential books and publications. Different names refer to essentially the same economical period in time of post-industrial era depending on witch aspects are stressed. In one of the most influential of those books Carl Shapiro and Hal Varian "Information Rules" [Shapiro and Varian, 1999] call it *The Network Economy*. Don Tapscott coined the term *The Digital Economy* to reflect that the goods are digital [Tapscott, 1999]. *The Information Economy* and *The Knowledge Economy* are also used by many to put the focus on the information goods and knowledge as the main source of value. Gosh Shikhar calls it *The Internet Economy* stressing the importance of the Internet [Shikhar, 1998]. Some people simply took the fact that the new situation was nothing like it was

before and call it *The New Economy*, including the founding-editor of Wired magazine Kevin Kelly - "New Rules for the New Economy" [Kelly, 1999]. In fact it doesn't really matter how we call it, as long as we all know what we are talking about.

How new is the new economy anyway? Shapiro and Varian argue that the underlying economic principles are not new - they just need to be applied differently with regards to the changed situation. As they say "Technology changes. Economic laws do not" [Shapiro and Varian, 1999]. Going even further into the past, over 200 years ago, the father of classical economics Adam Smith observed that "the division of labor is limited by the scope of the market" [Smith, 1776] because of the need to amortize fixed costs. What has changed now is the technology and thus, both the size of the addressable market and the relevant fixed costs of production and distribution. Internet and other inventions expand the scope of the market multifold and in most cases even make it a global market with marginal distribution costs close to zero.

Even though the economic laws remain the same, the application of those laws are very different from what we had in a pre-internet era. In that respect everything indeed does change. Everything from pricing strategies, lock-in consequences all the way to intellectual property rights and standards. And the main reason why everything has a new perspective has to do with the the properties of digital information as a new form of goods.

One of the key observations was that digital goods are very different from physical ones. Information has the following two important characteristics. First, it is costly to produce, but cheap to reproduce. That is, information is a good with high fixed costs but low marginal costs. "Information is thus a good with substantial (supply-side) economies of scale" [Varian, 2003]. Knowing that the marginal costs of production is one of a main sources of scarcity in the physical world, this open up lots of new possibilities for information goods and directly influences the amount of goods available and their pricing, which is based on value of a good rather than the cost of production. When economists say that information good is cheap to reproduce, good examples are information goods that still have some kind of physical form - books, movies, CDs, DVDs, etc. If we talk about pure information goods, meaning that they can are just a string of bits, then the cost of reproduction not only is smaller but actually drops to all the way to zero. Therefore pure information goods are "infinitely expansible when its quantity can be made arbitrarily large arbitrarily quickly at no cost" [Quah, 2003].

Second characteristic is that information is an experience good. It means that we have to experience it in order to know it's value. On one hand this leads to lots of free samples so that people could try things out, on the second

hand it also means the increased value of recommendations, since we can not try everything ourselves, we can be guided by other people's experience. In other words, information goods exhibit network externalities or network effects [Shapiro and Varian, 1999]

One interesting phenomena of the new economy when it comes down to media is the blurring line between professionals and amateurs. Digital photo and video cameras, coupled with personal computers and simple editing software makes everyone a producer. In this case we don't even need to talk about the price of good because almost all of the user generated media is for free. Coming from an industrial economical age we might say that nothing is for free and that everything that is produced is done so in order to make money, and in the case of broadcast media, we might even add that it needs to make big money. This is not the case in a user driven media world. There are plenty of other reasons for someone to produce rather that to earn a living. Self promotion, self expression, sharing or simply making your own mark in this world and just a few of the reasons. All this leads to ever increasing amounts of media available both on the professional and amateur side.

But all this media means nothing if it simply sits in someone's computer, it need to be accessible for other people. This is where the Internet comes in. Internet penetration is constantly increasing and the advances in data transmission technologies makes it work faster and thus enable rich media delivery. Internet, coupled with decreasing prices of storage on the computer side, is becoming a powerful media distribution platform.

Broadcast is still, and always will be, the most efficient platform for mass delivery, but the user needs, in the new economy setting, have changed. The question is, how much of the "mass delivery" we still need. While providing low prices broadcast inevitably narrows down the selection simply because of it's physical cappabilities. Nowadays we have witnessed that "increased ability to search for and find a broader variety of titles is 5-7 times more important than the lower prices" [Brynjolfsson et al., 2003].

To conclude, the two main factors that have fundamentally changed media world are the developments of computers (content production platform) and the Internet (content distribution platform). We are now entering the era of effectively infinite shelf space. Two of the main scarcity functions of traditional economics (the marginal costs of manufacturing and the distribution) are getting closer and closer to zero [Anderson, 2006].

1.1.3 The Long Tail of media

When Criss Anderson published the "The Long Tail" article in 2004 in a magazine Wired, it quickly became one of the most popular articles the magazine has ever had. In the article Anderson described the effects of the long tail on current and future business models. Anderson built his work on the earlier research done by Erik Brynjolfsson, Yu Hu, and Michael D. Smith, who were the first ones to use a two-diomensional graph to describe the relationship between Amazon sales and Amazon sales ranking. They have found that a large proportion of Amazon.com's book sales come from obscure books that are not available in brick-and-mortar stores [Brynjolfsson et al., 2006]. For years we were using Pareto principle, also known as 20/80 rule, which states that 80% of the income comes from 20% of the products. But what Brynjolfsson's team has shown was quite the opposite. The main reason for this is that once we eliminate scarcity the media consumption pattern no longer follows the Pareto principle but takes a shape of statistical distribution called *the long tail*, also known as heavy tails or powerlaw tails. In figure 1.1 you see an example of such distribution. It shows Amazon's weekly sales data. We can see that around 130,000 books sell at least one copy per week. Only the biggest of the physical retailers have the scale to have such number of titles in their store. 130,000 sounds like a lot of books, but there are around 3,700,000 published book titles in the world today. Where are the the other 97% of the books? Are they failures not worth mentioning? If a book fails to get into a top 130,000 list then how good can it possibly be and who could ever be interested in buying it? This was the way we used to look at it. Constrained by the limited shelf space we were forced to to look at it this way. But the true demand curve showed that people do not stop buying books beyond 100 thousand, 200 thousand or even 3 million. The data from Amazon have shown that no mater how much variety they offered to the customers almost all of the books get sold. Not by big numbers, maybe one per month, or even one per year, but they all sell. Traditional retailer does not have a luxury to keep such a variety since every inch of shelf space costs money, but when that space doesn't cost anything, suddenly we can look at those infrequent sellers again, and they begin to have value.

Statistically speaking long tail distribution is not a new thing. In fact every industry can in principal have a long tail distribution of sales if it meets three requirements. According to Anderson there are three factors that need to be in place in order for a distribution to take shape of a long tail:

- variety
- inequality
- network effects

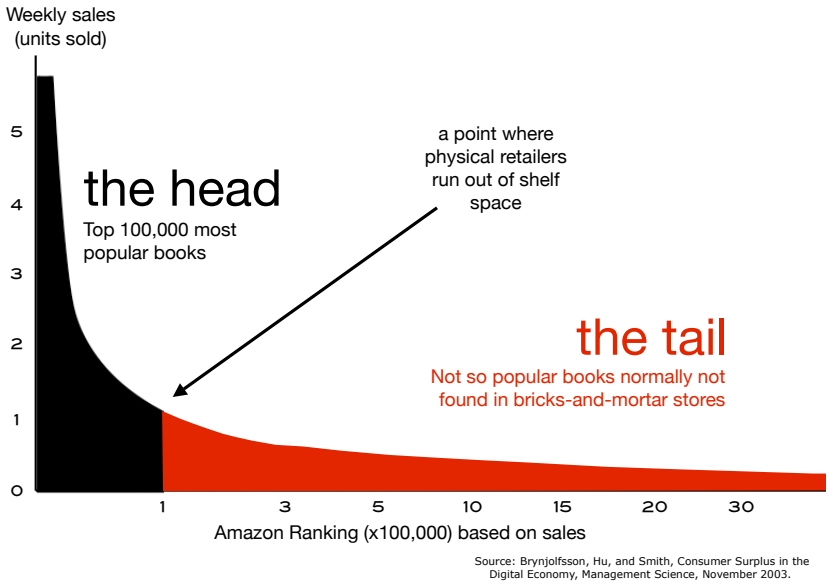


Figure 1.1: The long tail of Amazon's book sales distribution

First there need to be a great variety of different products to begin with. Then some of those products have to have higher quality than others, so that some would be better and some would be worse. Finally there needs to be some form of network effects (word of mouth, reputation, etc) which will amplify the differences in quality.

From those three requirements we can see that todays media meets all of them perfectly. From all the goods in the world the media has the greatest variety, and it is expanding all the time. There are definitely differences of quality in media going all the way from an award wining movies and greatest books of all times, to amateur videos taken with a mobile phone in an underground concert of an unknown band. And finally there are network effects (word of mouth, reviews, etc.) that help us to find quality media.

The research by Anderson and Brynjolfsson shows that once being exposed to a greater variety people start wondering away from the mainstream hits and dive more and more into the niches. That mass of niches has always existed, but they were not easily accessible. Now as the cost of reaching them is finally falling down it is suddenly becoming a cultural and economic force.

Their research has also showed that all those niches when aggregated, can make up a significant market. No wonder that some of the most successful companies

of the Internet age were the ones that saw that value in the tail and managed to utilize it. Such companies are: Amazon (books, DVDs, CDs, etc.), Google and Yahoo (long tail of search and advertising), YouTube (videos), Rhapsody and Last.fm (music), Netflix (DVD rentals), MobiTV (multimedia content for mobile devices), Audible.com (audiobooks) and lots of others.

Former music industry consultant Kevin Laws puts it this way: "The biggest money is in the smallest sales." To back up his words here are some sales data from several of the leading companies from the media industry: 25% of Amazon's book sales come from outside its top 130,000 titles; Rhapsody streams more songs each month beyond its top 10,000 than it does its top 10,000 and 40% of Rhapsody's sales come from beyond it's top 55,000; The average Blockbuster carries fewer than 3,000 DVDs. Yet 21% of Netflix rentals are outside its top 3,000 titles. [Anderson, 2006]

This is the difference between push and pull, between broadcast and personalized taste.

1.1.4 The rules of the Long Tail

The Long Tail theory essentially boils down to two simple rules.

- make all media available
- help customers find media

First we must make all the content available for the users to reach. This may sound simple, but in fact there are two main obstacles. First of all, some media that has been created years ago may be already lost and since it was never digitalized it can not be recovered, or even though it may still be present, but because of it's analogue nature is can not be distributed on the Internet, and since in most cases such media never made it big it may be economically not worthwhile to digitalize it. The second problem is the legal issues. If the first one targets mainly old obscure media, legal issues affect even newly produced popular content. There are numerous examples of quality content that is not available or needs to be modified due to the legal rights. *WKRP in Cincinnati* was one of the most popular television shows of the late '70s and early '80s, but it is unlikely ever to be released on DVD because of high music-licensing costs [Dean, 2005]. Another example could be the 1987 *Married with Children* - the second-longest-lasting sitcom on the FOX network (second only to The Simpsons) nominated for 7 Golden Globes. Despite it's world popularity and

demand it could not be released on DVD because a Columbia Tri-Star was unable to obtain the rights to the theme song (Frank Sinatra *Love and Marriage*). It has eventually led to re-editing the episodes and thus disappointing millions of fans. I will not go into any legal issues in this thesis and will concentrate on the second rule of the Long Tail - helping users to find the content they want.

From the examples of Amazon, Netflix, Rhapsody, iTunes, and many others who have already tapped into the long tail of media, it is clear that more available content (or in general, more information) leads to greater sales for the suppliers and better experience for the users. It sounds like a win-win situation - all we need to do it to simply throw all the media out there in the Internet and see what happens. Most likely nothing would happen. Most of the researchers who have worked with the information economy eventually make one simple but crucial conclusion - the main scarcity is the human information processing capabilities. We are not digital creatures and unfortunately our brain can not process huge amounts of data efficiently. Economist Herbert A. Simon puts it: "...in an information-rich world, the wealth of information means a death of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it." [Simon, 1971]. Therefore it is our attention and time that are the main bottlenecks in the new economy.

Even though the numbers suggest that in some cases there may be more money in the tail than in the head portion of the curve, both ends of the tail are still needed. Big hits (broadcast) still matter - they act as an attractor. If we have the products only at the head, we will inevitably run into a situation where the customers want more and we can not offer it. On the other hand, if we have just the products at the tail, we will find that customers have no idea where to start since everything is unfamiliar for them. The importance of having both the head and the tail is that then we can use the the mainstream hits as a point of entry and use them as the attractors just to get people in. But once they are inside, recommendations can be used to guide customers further into the unknown and vast spaces of obscure media. This observation leads to the second rule of the Long Tail economics - helping customers to find new unknown media through recommendations because by simply making content available will not drive the demand down the tail.

1.1.5 Search is the key

In todays media world we really do not lack variety. What we do lack is the ability to efficiently find media we like. And if we find what we like, we have no idea of what else might be out there that we may like even more. How good is the unlimited variety if we can not find anything? Without the search the long tail of media is just a huge pile of content with little value. Numerous studies report that the search functionality alone changes customer's consumption patterns tremendously. Blockbuster has reported [Anderson, 2006] that about 90 percent of the movies they rent are new theatrical releases. In the case of Netflix the new releases are only about 30% and about 70% is back catalogue. Anderson argues that it is not because they have different subscriber base, but it is because of the search ability that Netflix offers to it's customers creating demand for the niche content. And they do it algorithmically with recommendations and ratings. In fact, 60% of Netflix rentals come from recommendations.

Another study in 2005 by MIT lead by Brynjolfsson looked at women's clothing retailer. Customers who shopped both online and using the catalogue tended to go further down the tail online exploring the niches more. The bottom 80% accounted for 15.7% for catalogue sales, but the same 80% acounted 28.8% online [Brynjolfsson et al., 2006]. The reason for this is the search, concludes Brynjolfsson.

The following quote serves as a good illustration of how the search is influencing our choices and thus changing everything else (taken from Frog Design consultancy company) – "We are leaving the information age and entering recommendation age. Today information is ridiculously easy to get; you practically trip over it on the street. Information gathering is no longer the issue - making smart decisions based on the information is now the trick... recommendations serve as a shortcuts through the thicket of information, just as my wine shop owner shorcuts me to obscure french wines to enjoy with my pasta."

We used to have filters of media all the time. They are the ones that filter media before it even gets out (pre-filters). Those were the editors of magazines, managers of broadcast TV stations and music record studios. They were acting as a gatekeepers controlling which media gets out and which doesn't. Another way of thinking would be to let everything get out there and only then filter media through recommendations and positive feedback. This is exactly how it works in the Internet. "Recommender systems have the potential to automate word of mouth, speeding the discovery and diffusion of new goods" [Resnick and Varian, 1997].

1.1.6 Media recommendation

Since the main technological challenge in the Long Tail economy is how to find relevant media, this thesis aims to contribute in this area. There are two main ways to find interesting media. If we more or less know what we are looking for then we can use the search engines to get us to a desired content. But judging from the vastness of the media available today, we can not possibly have any idea of what else is out there and if we don't know that means we can not even formulate the search query.

The second approach is to use the collected knowledge of other people by letting their recommendations guide us to new things. The market success of the recommender systems and numerous scientific conferences on this topic serves as a good indicator that this is a very important issue. The long tail of digital media would not exist without recommendations. Therefore in this thesis the main goal is to improve the recommendations of media by combining knowledge of various fields. The rest of this chapter is meant to quickly introduce those fields and give an overview of how they all fit together before going into any detail in any of them since it is done in the remaining chapters of the thesis.

All recommendation techniques are based on the ability to find similarity, in one way or another. There are two popular approaches – content based filtering is based on finding similar content, while collaborative filtering starts by finding similar users (based on the content they consumed, therefore is leads back to ability to find similar content). Since being able to find similarity of two media items is the key here, how do we go about it? First of all it depends on what we know about those items (so that we could compare them), where metadata is the term used to describe characteristics of the content. Categorization of media can either be done using predefined "top-down" categories or based on "bottom-up" media features.

There are many different supervised metadata standards for all kinds of media: text (Dublin Core), images (Exif, DIG35, XMP, Z39.87), audio(ID3, OGG Vorbis, APE), video(MPEG7, TV-Anytime), news(IPTC), etc. Most of them are based on keywords, some support short textual descriptions, and a few also use predefined classification terms. TV-Anytime can be viewed as a subset of the MPEG7 standard that supports all of those ways to describe media. In the thesis the part of the empirical data is BBC programs annotated using TV-Anytime. "Bottom-up" approach could be illustrated by the folksonomies. Social networks like Last.fm or Flickr are all based on user generated tags describing the content using user generated tags. While being not professionally annotated those tags reflect the audience and show how people perceive certain media items. User generated tags are especially helpful to indicate an emotional response to the

media (happy, sad, violent, angry, etc.) complementary to existing metadata standards.

Most media has an informative value (news, TV content) or emotional value (music, movies), or in most general case - both. The informative value of media has become a large research area and a number of methods have been developed, ranging from signal analysis (looking for certain features in media) to information retrieval analysis of keywords and categories assigned to the content. Emotional side of media is harder to extract because it is much more complex to describe in the first place as it additionally involves cognitive and psychological aspects. Synopsis, or any textual description, is a good indicator about what the content is like. In the thesis synopsis and song lyrics are used for automatically extracting the emotional context of media.

1.2 Research question

As stated before, the main technological challenge in the Long Tail markets is how to find relevant media. Based on the state of the art of the research areas that deal with finding media – information retrieval, recommender systems and media annotation – the main problem is how to automatically find relevant media in when we have so many options to choose from. This is a very top level question. During the research process, the question gets more and more precise every time we find out more information about the various components of personalization.

The final version of the question sounds like this:

> *How to improve media recommendation by utilizing the emotional components of media, and how to extract the emotional value automatically from the metadata?*

1.3 Structure and outline

Knowing the research question of the thesis and the empirical data, certain theories and methodologies can be applied. This section briefly presents those methodologies and outlines the plan of how to answer the research question (see Figure 1.2).

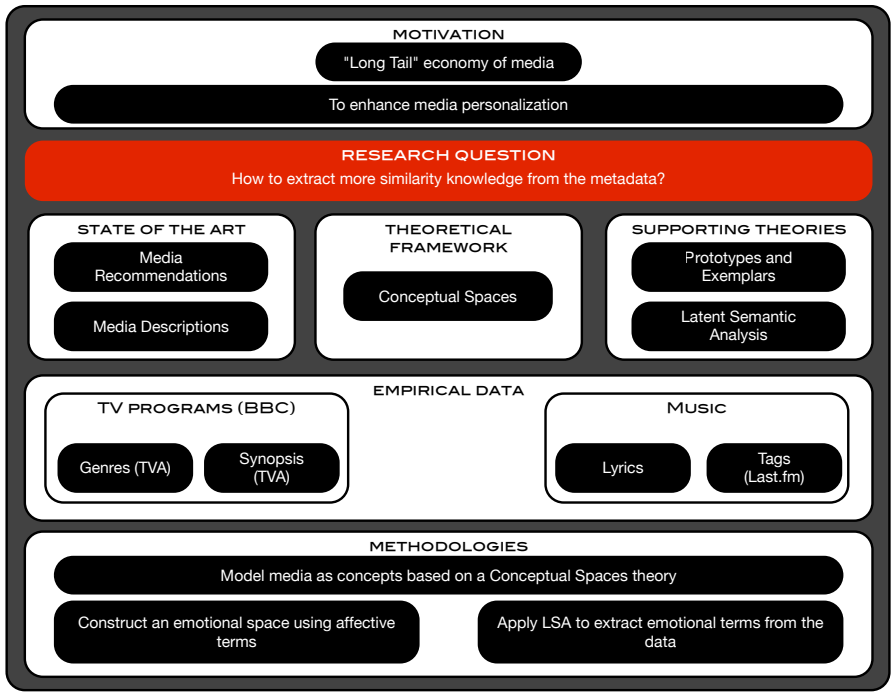


Figure 1.2: The schematic view of the structure of the thesis for approaching a media personalization problem

This introduction *Chapter 1* has presented a motivation for the research by presenting the current economic phenomena in the media world today called – the Long Tail – and pointed out that in todays world the main problem is how to find interesting high quality media in the ocean of alternatives. Therefore we start with the main research problem – how to find relevant media in vast amount of alternatives. The actual research question became clear only when a number of media personalization components were analyzed and it is presented in the previous page.

In order to solve the research problem first I present the structure of the personalization system outlining all of the relevant components (see Figure 2.1). There are two essential parts in the picture. First one is the recommendation algorithms, explaining how the certain items are being selected and why, while the second one looks into the items themselves and represents the media metadata field. These two areas are analyzed as the state of the art in the *Chapters 2 and 3* respectively showing what can be done at the moment and where are the main bottlenecks. *Chapter 2* presents the current research in the area of the

recommendation systems, explaining their types, methods and approaches. The notion of "similarity" appears to be very central to the whole idea of recommendations, therefore *Chapter 2* also talks about how different recommendation methods approach the similarity problem and how different item-item or user-user similarities are calculated using standard techniques.

Chapter 3 continues the "similarity" idea and explores the metadata pillar analyzing the different kinds of information that makes up the metadata and what kind of similarity knowledge each metadata type can provide. It appeared that the most potential lies in the unstructured metadata, which is primarily targeted to humans rather than machines. Another point is that emotional value appears to be very important when talking about video, and especially audio, content. One of the main hypothesis made at this point was that the emotional metadata is capable of identifying content items which might be perceived as similar and thus increase the number of relevant recommendations by capturing features across the traditional divide of categories. Such hypothesis was tested in *Chapter 5* where I presented the first empirical analysis of the TV personalization based on genre information. The first case showed the limitations of the approach to the similarity estimation based on the feature overlaps. But even as limited as it is, it has also highlighted the emotional genres being able to cross the traditional genre categorization boundaries.

At this point it became clear that in order to move further we need really understand what is the "similarity" on a very fundamental level, and what are the ways to approach it. This takes us to the the cognitive side of the science (*Chapter 4*) where different similarity and categorization theories are discussed. The two main categorization models are presented in this chapter, – Prototype and Exemplar models – eventually leading to the main theoretical framework – the theory of Conceptual Spaces by Peter Gärdenfors. I hypothesize that the meaning of media can be represented using Gärdenfors Conceptual Spaces theory. Conceptual Spaces is applied in this thesis to analyze media in terms of its dimensions and knowledge domains, which in return defines properties and concepts. It turns out that one of the most important domains in terms of media is emotional one, especially when we talk about such media as music. Therefore the main focus in the thesis is how to extract such emotional information from media, and how to use it in order to model the concepts of media items. This point of view is carried over through the rest of the thesis resulting in one of the main scientific contributions – application of Gärdenfors theory to media by using emotions as a domain.

In the *Chapter 6* I make a hypothesis that it is possible to automatically extract emotional information from the unstructured metadata since it reflects the inner structure of the media. As a proof I propose a novel method to extract emotional information from text using Latent Semantic Analysis (LSA). I

present two separate cases to illustrate the similarity knowledge extraction from the metadata, where each case extracts emotional value using the metadata that represents different abstraction levels – *synopsis* and *lyrics*. The emotional value is extracted by first creating a conceptual space for emotions based on the studies of semantic differential which divides the emotional space along two dimensions – arousal and valence. Then certain regions in the space are selected to serve as emotional markers – the affective terms. After that LSA is used to calculate the cosine similarity between the text (synopsis or lyrics metadata) and each of the 12 affective terms. As a result we get emotional patterns, which we can then use to compare media based on emotions.

By being able to compare media items on the basis of their emotional patterns, we add a new level to how we can evaluate the similarity between two media items. Which in return might improve media recommendation since it provides a novel approach to recommendation that goes beyond traditional genre boundaries, and thereby improves media personalization.

Thesis finishes with *Chapter 7* where I present the discussion of the results, my overall conclusions and a few pointers for the future research.

The appendixes A, B, C, D and E contain five publications by the author addressing each of the three media personalization cases – TVA Genres, TVA Synopsis and Music lyrics from a number of different perspectives. The first two are published as Springer LNCS book chapters, the third of is a published in the conference proceedings and the final two as journal articles.

1.4 Scientific contributions

This thesis aims at proposing a novel approach to automatically extract the emotional aspects of media from the unstructured metadata. Theoretically the thesis spans over many different areas and uses different methodologies. The main contributions are the following:

- Analyze the problem of recommendation, by evaluating different approaches currently available in the market.
- Identify the metadata elements that are the most important for the media personalization.
- Present the cognitive science theories and models that are relevant to the media personalization problem.

- Apply Gärdenfors' theory of Conceptual Spaces to emotions, by using emotional domain to model the concept of media.
- Propose a model for constructing the emotional space, by selecting 12 terms to serve as emotional buoys or markers.
- Propose a novel approach for extracting emotional terms from either *synopsis* or *lyrics* metadata using Latent Semantic Analysis.

Chapter 2

Media Recommendations

One of the main differences between the media world of the past and the one of today is the unprecedented amounts of content available. Advances in technology have eliminated the scarcity of storage and distribution. This in turn has resulted in the explosive growth of niche media that was simply not feasible to produce in the past because it would not generate enough revenue to cover the production costs. What we see today is the media world unfiltered by scarcity. This is clearly a positive thing and it has never been a better time to be either an artist or a consumer, but it raises several problems as well. The main problem is how to find interesting high quality media in this ocean of alternatives.

There is no doubt that we are all different. It is especially true when it comes down to our taste in music, books, movies or any other forms of media. And yet we all have access to the same pool of information - the Internet. In other words, on one side we have all the media in the world - a massive impersonal database, and on the other side we have an individual with unique taste and unique needs. Therefore the main challenge is to somehow be able to build an custom tailored interface between the individual and the content. This process is also known as *personalization*.

This chapter presents the state of the art in the area of media personalization presenting the main ideas, methods and techniques.

2.1 Media personalization

Lets begin by defining what is *media* and what means to *personalize* media. Media is a broad term that can mean a number of things. We have *print media*, *mass media*, we even have the *new media*, the list literally just goes on. In this thesis media refers to digital content (pictures, audio, video, text) available on any digital delivery channel, like the the Internet or digital TV. In all cases there is a big difference between how humans perceive media and how computers see it. That difference is called the semantic gap, and the more complex the media, the bigger the gap. The most challenging types of media are the audio and video because there is a huge difference between their low level features (the way computers see it) and semantic meaning (the way humans see it). Therefore in this thesis the main focus is to improve audio and video media personalization by introducing novel approaches and contributing to the field of media retrieval.

Personalization as a process refers to tailoring goods to fit individual needs. It is still a young and developing field, therefore there still exist different points of view on what personalization is, expressed by academics and practitioners. One of the "official" definitions of personalization is the following:

> *"Personalization is the combined use of technology and customer information to tailor electronic commerce interactions between a business and each individual customer. Using information either previously obtained or provided in real-time about the customer and other customers, the exchange between the parties is altered to fit that customer's stated needs so that the transaction requires less time and delivers a product best suited to that customer". (www.personalization.com)*

The personalization objectives usually are multifaceted. They may range from simply improving the consumer's browsing and shopping experience (e.g., by presenting only the content that is relevant to the consumer) to much more complex objectives, such as building long-term relationships with consumers, improving consumer loyalty, and generating a measurable value for the company [Adomavicius and Tuzhilin, 2005a]. In the case of media, personalization is becoming a necessity rather than an a luxury feature. The availability of massive amounts of content are worthless if they remain hidden in the long tail of media.

Basic types of content personalization fall into two categories: rules based and information driven [Ha, 2006]. Rules-based personalization delivers content on the basis of decision rules made from user profiles. In most cases it tends to be more static, with rules established in advance, for example "if this then

that". Information-driven personalization can be relatively dynamic and usually adopts one or a combination of three types of filtering techniques: content based, collaborative and hybrid approaches.

Quite often terms personalization and customization are confused. They both are very similar in terms of their goals, but differ in the way the user is involved. Customization is a process where the user is in a total control and is given the capability to modify the product - to customize it. It implies the manual involvement of the user, rather than being done automatically. The control of the look and/or content is explicit and is user-driven, i.e. the user is involved actively in the process and has control. In personalization, on the other hand, the user is seen as being passive, or at least somewhat less in control. Customization is also called user-controlled personalization [Ha, 2006].

Another important aspect is to define where personalization happens. Personalization can be performed on several different levels. Different researchers distinguish from two to five, or even more different levels of personalization (product, price, service, interface, etc). When we talk about physical goods, the object of personalization quite usually is the physical product itself - electronic devices, jewelry, clothes, etc. Every product has a number of features that can be modified or adjusted to suit each individual best. In such personalization the main challenge is to identify the features of the certain product that customers value most and then provide multiple alternatives for those features [Tapscott, 1999]. Of course in the physical world there are always limitations as to which aspects of the product can be modified in an economically feasible way.

In the case of digital media usually it is not the product itself that is the focus of personalization. That means that the goal is not to modify the media itself, but instead to address the main problem in the media world - search and retrieval of relevant content, therefore media personalization usually deals with personalizing the interface between the user and media databases. Media personalization can be defined as:

> *A process where the vast amounts of media are taken and automatically narrowed down to a limited set of items that fits best for our needs in a given situation.*

From this very top level definition, we can already see the four main parts of the media personalization system: media itself, our needs (user preferences), the situation (context) and a mechanism to match media to users in the light of the context (Figure 2.1). Media personalization process is all about connecting the users with the content. All the publicly available media is located either on

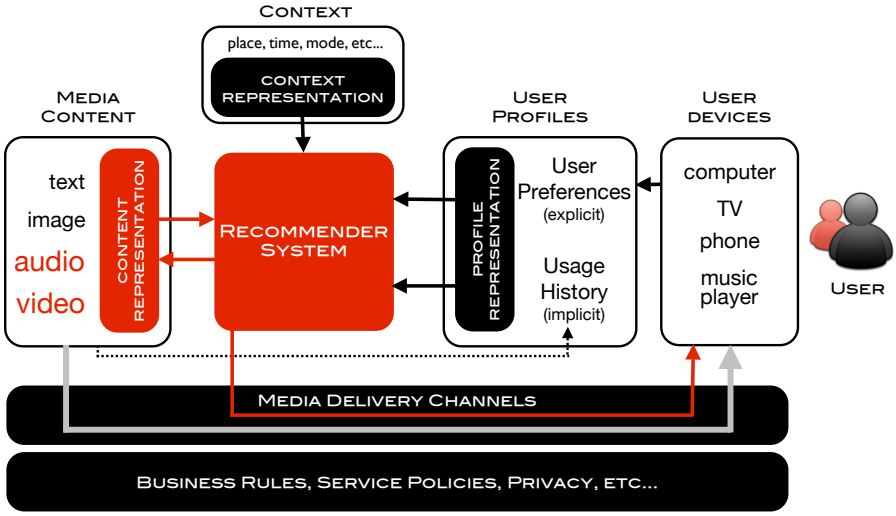


Figure 2.1: The different elements of a media personalization system.

servers or personal computers and is accessible over the Internet, mobile networks or digital broadcast (TV and radio). Depending on the delivery platform users are equipped with variety of devices - usually a computer, mobile phone, TV and radio device. Internet is clearly the dominant platform for digital media and due to it's on-demand nature it can store literally unlimited amounts of media, whereas in broadcast there are very clear limitations to how much media can be transmitted at a time. Every device has an interface which enables users to navigate and interact with the content. Every service also has an interface that depends on the type of the service and on the device. Examples of such interfaces would be Electronic Program Guide (EPG) on our TV and radio, or personal homepage at Amazon.com, Last.fm or YouTube. In some cases in the past it was possible to simply put everything that was available in the form of an organized list and present it to the user, a good example would be an old paper TV Guide that had every single program that is visible in the given broadcast area. It is simply not possible to do this any more because of the several reasons. First, there is way too much content available - one could spend a lifetime just looking through the list. Secondly, the amount of media is constantly growing and new items are appearing all the time. It means that by the time we finish reading the program list, it is not the same list anymore, since big part of it has already changed. And finally even if we could look through the list of content manually (imagine a list of iMDB's top 250 movies of all times) and if we assume that the list is relatively static, we still have a problem deciding witch movie to pick because we either do not have enough information

about each movie to make a decision, or we have loads of information about each movie and then we simply can not process this information manually in an efficient manner. There is only one way to go about it - filtering of media needs to be done automatically.

If we agree that we have to let machines to personalize our interface into media, then the first problem is how to make machines understand what the media is like, to understand what the user likes and to correctly evaluate current context. Second problem is once machines understand to a certain degree content, users and context, how to use all that input to produce a personalized output.

First problem is solved by describing resources (media, users, context, etc) in a machine understandable way. The process is called annotation if we talk about content side, or building user profiles if we have user side in mind. In other words, all data that comes as input into the system need to be represented in order to enable computers to understand it. *Chapter 3* presents the state of the art in the field of media representations and discusses the importance of different types of content annotations.

When it comes down to understanding the users, there are two ways to find out what user likes. One way is to let the user state preferences explicitly. This works in situations where user knows exactly what she wants and why she wants it. But the more complex the product, the harder it is to specify why we like what we like. This is exactly the case in media. For most people it is very hard to describe what kind of movies or music one likes. Once asked to do so, people usually start giving examples of something they like instead of trying to describe the media they like by using features. And in the case of media that works very well. If I need to tell somebody what kind of music I prefer, it is most informative to simply give the list (lets say top 50) of artists that I listen to. This kind of information is called *usage history* and in most cases serves as the main knowledge resource used to get into the user's mind. It is easy to gather such information, all that is needed is that records are being kept of everything that user does - all the media user buys, consumes, views, rates. This is already done for every media service that requires users to identify themselves by logging in (Amazon.com, Last.fm, etc.). Important point to mention is that usage history mirrors the media content side. It means that the more we know about the content - the more we can tell about the user who uses that content. This shifts the complexity of representing resources to the content side. In this thesis one of the assumptions is that we can sufficiently describe the user just by using usage history. Putting most of the user profile weight on usage history instead of explicit preferences is not a novel idea and is supported by a number of researchers [Konstan et al., 1997, Goldberg et al., 1992, Nichols, 1997, Rucker and Polanco, 1997] and also form the basis for collaborative filtering recommendations.

Context is all the information that describes the current situation that user is in: location, time, motion, mode, who the user is with, etc. Context gives an extra input into the system and sometimes can play a major role in media personalization process. For instance, knowing if the user is sitting at her desk, walking, or driving a car is crucial in order to decide what media format is preferable in those situations: long video, short video, audio-only, etc. There are other situations where user location or time of a day is a major parameter. Context information is gathered by sensors embedded into user devices or/and with the help of the network that is serving the device. In the case of media today context mostly affects the presentation format of the content rather than influences the selection of the content itself (in principal context can and does influence the selection of content, but the process of extracting and interpreting context today is very complicated and therefore in most cases is not used yet). Therefore another assumption in this thesis is that we have no reliable context information available and will not take context into consideration while personalizing media.

There are a number of other factors that may have major impact on the way personalization works. These are the factors driven by economic and regulatory forces. They deal with issues of what is and is not allowed, and how to influence personalization process to maximize profits. Another big questions is privacy. Even though these are the areas that one would need to address before implementing personalization in the real world, these challenges do not play a significant role in the actual content selection process and therefore will be out of focus in this thesis.

So far we have the users with their devices accessing various forms of media over various delivery channels. Both media resources and users are represented in a form understandable to computers in order to enable automatic processing. At the very center of the personalization system lies the recommender engine. It takes media descriptions, user profiles and contextual information (if available) and produces personalized content recommendations for each user. As we will see later in the thesis, it is not the recommender part itself that is the bottle neck in the personalization system but rather everything that comes in as an input. Nevertheless it is crucial to understand exactly how recommendation process works because this gives a good idea on what all the other parts have to look like in order to achieve best overall results.

2.2 Recommender systems

Recommender systems emerged as an independent research area in the mid-1990s to address the problem of information overload. The main goal of a recommender system is to "produce individualized recommendations as output or have the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options" [Burke, 2002]. Ever since there has been much work done in both industry and academia developing new recommendation techniques and approaches. Nevertheless after more than a decade of intense research the interest in this area remains very high.

There have been many implementations of recommendation systems in various industries. Even though this field as we know it has emerged a bit more than a decade ago, the first ideas and implementations appeared as far as 30 years back. Library science is only one of the numerous disciplines that modern recommendation systems build on, therefore there is no surprise that the one of the earliest implementations come from that particular area. *Grundy* system [Rich, 1979] is considered to be the first recommender system, which proposed to use stereotypes to build individual user models and use them to recommend relevant books to each user. Few decades latter, the *Tapestry* system [Goldberg et al., 1992], which is considered the first collaborative filtering system, was introduced. Both similar user neighborhood building and filtering was performed manually. On top of that it scaled badly and thus was not suitable for the mass public. Few years latter the first recommendation systems for the mass public appeared. *GroupLens*, a research group form the University of Minnesota, applied recommendations to the areas of news articles (*UseNet*) [Resnick et al., 1994, Konstan et al., 1997]. One of the first commercial companies to realize the potential of niche media *Amazon.com* started using recommendations first for books, latter for many other types of goods [Linden et al., 2003]. *Ringo* and its successor *Firefly* applied recommendations to the music domain [Shardanand and Maes, 1995]. *Video Recommender* [Hill et al., 1995] and *MovieLens* implemented recommendation techniques to help the users to discover new movies. Recommending TV programs was also explored, with notable publications by [Gutta et al., 2000, Miller et al., 2003, Yu and Zhou, 2004, Sullivan et al., 2004].

In the last few years, with the explosive growth of the Web 2.0 and the appearance of all kinds of social networks recommendation systems have been implemented in all possible domains of media. A few notable examples include YouTube, Last.fm, iTunes, Netflix etc. The main challenge still remains how to make recommendations more useful and the limitations of such systems have been anything but reached.

According to one of the leading researchers in the field of recommender systems Gediminas Adomavicius (University of Minnesota) "the recommendation problem is reduced to the problem of estimating ratings for the items that have not yet been seen by a user" [Adomavicius and Tuzhilin, 2005b]. This definition highlights one of the main requirements for what we consider a good recommendation - it needs to be novel. This is exactly what separates recommendation and search paradigms. In search we start by formulating a query and then the task for a search engine is to find content item with the highest possible match to fit the query. Since in most cases queries tend to be very abstract we get thousands of results ranked by how much, according to a search engine, they fit our query. When we search for something we know to some extent what we are looking for, otherwise we would not be able to formulate the search query.

Recommendations take fundamentally different approach and instead of aiming to answer our search query as precise as it can, it suggests novel items that it thinks (based on a number of techniques) we might like. Novelty is very important here because it helps us to drift away from the mainstream media and dive into the niches that we otherwise would not be even aware of. From the economical point of view recommendation have already proven to be a successful way to sell media and in many cases (for example Netflix or Amazon.com) is responsible for significant share of revenue. Recommendation and search are often coupled together where the search functionality is helping users to find the media that they know and then taking this as a point of entry to recommend other relevant and novel items.

Before analyzing the actual recommendation algorithms, it is important to understand what data goes into the system as an input, and what output we are aiming at in the end. Since the main goal to recommend new products to the customers, two primary sources of input are the *user space* and the *item space* [Popesculand et al., 2001]. Item space consists of n available items i where each item is identified with a unique id which is necessary to tell items apart. Item j will be referred as the item for which prediction is sought.

$$ItemSpace \qquad \mathcal{I} = \{i_1, i_2, \ldots, i_j, \ldots, i_n\}$$

Similarly user space consists of all the users in the system, u_a being called the active user for whom the predictions are calculated.

$$UserSpace \qquad \mathcal{U} = \{u_1, u_2, \ldots, u_a, \ldots, u_m\}$$

User and item spaces are enough to make recommendations even if we know nothing about neither of them, other than which user bought, consumed or rated which of the items. In this case all the meaning has to be derived from interactions between the users and the items. This connection is represented in the interaction matrix A consisting of elements $a_{i,j}$. But in most cases we

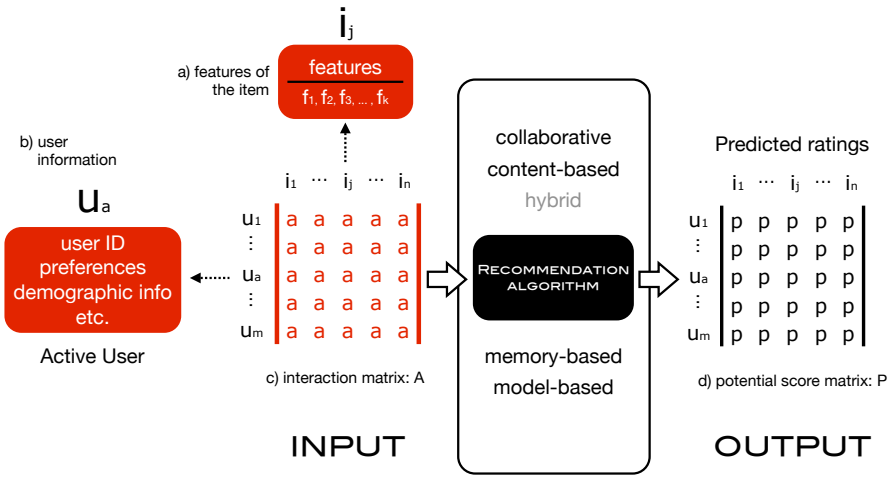


Figure 2.2: Recommendation process seen as an input and output.

do know something about either users or item. This knowledge can be then represented in appropriate form and used to make more meaningful recommendations. Every item can always be described with a set of features f which represent different qualities of the item. Even though this information is not always available it helps a lot if we have it because then we can base recommendations not only on the interaction matrix but also on features that the item has. Media features depend on the form of media itself (text, audio, video, ect.) and can be represented in many ways using a number of different standards. If we take all the features that items may have, they make up a feature space with k unique features.

$$FeatureSpace \qquad \mathcal{F} = \{f_1, f_2, \ldots, f_p, \ldots, f_k\}$$

To sum up, as an input into the recommendation system we may use an interaction matrix A generated as a result of interaction between user space and item space, and then use matrix A to infer the relationships between items and users. If we do not have interaction matrix, then we must get all knowledge about users and items through their features, which represent the qualities of an item (for example a genre of a movie) and the characteristics of a user (explicit and implicit preferences). In the ideal case, that actually is quite frequent we have both types of input at the same time (Figure 2.2). If we take the interaction matrix A and isolate a row for an active user u_a, we get a very sparse row with very few actual values, since there are way too many items for any user to have rated or consumed many of them [Linden et al., 2003]. It is nearly impossible even for the most active users to cover as little as 1% of content available (that

would equal to nearly 40 thousand books, over 4 thousand movies and several millions of songs). The goal of the recommender is then to fill that mostly empty row with predicted ratings trying to estimate how much user u_a will like every unseen item. To put it in another way, the recommendation system is trying to estimate how much the user will like items that she has not consumed yet. Such estimation can take form of prediction or recommendation [Sarwar et al., 2001].

- Prediction is a numerical value $P_{a,j}$, expressing the predicted likeness of item $i_j \notin I_{u_a}$ for the active user u_a. This predicted value is within the same scale as the opinion values provided by u_a
- Recommendation is a list of N items, $I_r \subset \mathcal{I}$, that the active user will like the most. Recommended list must be on the items not already purchased by the active user, i.e., $I_r \cap I_{u_a}$

More generally one can think of the recommendation process as a process to transform interaction matrix A coupled with a feature space $\mathcal{F}$ into a potential score matrix P (Figure 2.2). From this matrix we then can take the predicted rating value for the item i_j to get the prediction, or we can take the user u_a and collect top N items i for that user with the highest ratings ($\max R$) to get recommendation.

Recommendation techniques can be classified into a number of groups. Classification itself can be done according to different criteria. In terms of the input sources used, recommendation techniques are grouped into three different approaches [Adomavicius and Tuzhilin, 2005b]:

- collaborative filtering
- content-based filtering
- hybrid (combination of the other two)

Classical collaborative filtering (also called sociological filtering) completely ignores the features and solely relies on the interaction matrix A from which it then calculates predictions and recommendations. Content-based filtering (also known as cognitive filtering) relies only on the features of items and uses them as a fundament to predict which items will interest the user most. Both of those approaches have their own strengths and weaknesses so it is not surprising that researchers have found ways to combine them creating hybrid solutions to get the best out of both worlds.

Recommendation systems can also be classified into two big classes based on the actual recommendation process [Breese et al., 1998]:

- memory-based
- model-based

Memory-based recommendation techniques take user-item interaction matrix A, item features or user profiles exactly as they are, and then use various statistical techniques to calculate predictions and recommendations, whereas model-based techniques use the user-item interaction matrix to build a model first and then use it to calculate the actual recommendations.

These two classifications do not contradict each other since they use different criteria as a basis for classification. We could look at the whole recommendation systems taxonomy as a 3×2 matrix, where we have both classification systems put orthogonally to each other creating 6 unique classes. In the rest of the chapter different recommendation approaches are presented.

2.3 Memory-based recommendations

Memory-based recommendation techniques share the fact that they make recommendations based on the exact input without changing it. In the case of pure collaborative approach such input is the interaction matrix, while in the pure content-based approach the input are all the features that items and users have in their descriptions. Memory-based automatic recommendation techniques were the first ones to appear and are still very popular. The main idea is that the predictions are based on the readily accessible information using relatively simple rules and algorithms, this is why memory-based techniques are also called heuristic-based.

2.3.1 Collaborative filtering

One of the most popular and the most successful recommendation technologies to date is collaborative filtering [Resnick et al., 1994, Shardanand and Maes, 1995, Hill et al., 1995, Konstan et al., 1997]. Word "collaborative" here means that recommendations are based on the opinions of some other users than the active user u_a. Those opinions are expressed in the interaction matrix A, which therefore is the main input source for collaborative filtering.

Collaborative filtering techniques can be classified in a few ways. Just like all the other recommendation approaches it can be either memory-based or model-based, depending whether we use the interaction matrix directly or we build a model first. If we go into further classification, memory-based techniques can be user-based or item-based depending on how we calculate the neighborhood, wether we search for the most similar users or the most similar items (in both cases still based on the interaction matrix). Model-based techniques are divided depending what kind of machine learning technique they use to build models: clustering, bayesian networks, artificial neural networks, linear regression, probabilistic models, etc.

The idea of memory-based collaborative filtering is to suggest new items or to predict the utility of a certain item for a particular user based on the opinions of other like-minded users [Sarwar et al., 2001]. Like-minded users form *the neighborhood* for the active user u_a. One of the key features of this method is that the predictions themselves are then calculated based on the subset of users - the neighborhood - rather than the entire user base. This stresses the importance of the neighborhood selection. In the traditional collaborative filtering such neighborhood refer to the most similar users, for this reason this approach is also called user-based collaborative filtering. Another way to look at it is to form the neighborhood of the most similar items rather than users. This is known as item-based collaborative filtering. In principal user-based and item-based collaborative filtering methods are very similar and share their key principles. Both methods consists of two main steps:

- STEP 1: find neighborhood
- STEP 2: predict rating for unseen items

First we need to be able first to find the set of the most suitable users or items for the given task based on their similarities with either user u_a (user-based) or item i_j (item-based), and then the neighborhoods' ratings are combined to produce a prediction or top-N recommendation for the active user [Breese et al., 1998, Delgado and Ishii, 1999, Resnick et al., 1994]. In the nutshell, this is how memory-based-collaborative filtering works.

This class of recommendation techniques is the most popular in term of practical implementations due to the number of advantages. It is relatively easy to implement - all we need is the interaction matrix with the ratings expressing how much certain users liked certain items. This in return means that this approach can be applied to any domain since no explicit knowledge is needed neither about the users nor items. This is a significant advantage knowing how expensive it usually is to get quality metadata to describe items.

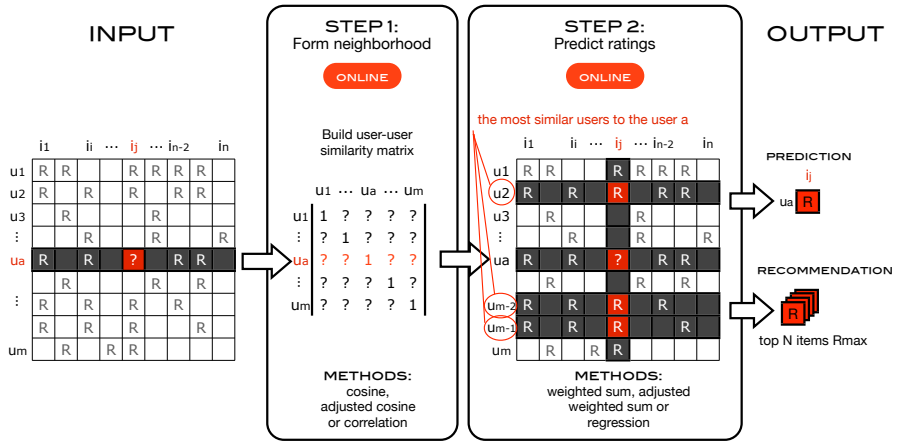


Figure 2.3: User-based collaborative filtering process.

Matrix A represents users opinions about the items. Opinions can be expressed explicitly given by the user as a rating score or can be implicitly derived from purchase records, by analyzing timing logs, by mining web hyperlinks and so on [Konstan et al., 1997]. If we use implicit knowledge to fill in the matrix, then the easiest way is to simply take purchase records and represent purchased items as 1 and the rest as 0. Then we can produce recommendations like the following "users who bought this also bought this". The main limitation of such approach is that the system does not know if the user actually liked the consumed item and if she liked it, then how much. More elaborate way is to let the user to express their opinions by letting them to rate items. This transfers into the matrix being filled with rating values instead of 1s and 0s and then matrix A can be called the *ratings matrix*.

User-based collaborative filtering is sometimes referred as a traditional collaborative filtering because it was the first one to appear and quickly became very popular (Tapestry, UseNet, Amazon.com, etc). The main assumption that collaborative filtering makes is the following - *users will like items that other similar users have liked in the past*. Before asking cognitive questions such as "what makes two things to be perceived as *similar*" lets state that every recommendation approach makes their own assumptions as to what makes two users or items similar. In the case of the traditional user-based collaborative filtering the assumption is that users are similar if they have a history of consuming the same items in the past. Or if we have ratings matrix then we can refine the definition by stating that similar users are the ones who have rated the same items in the same way in the past.

The filtering process is the following (Figure 2.3). We start with the interaction matrix A. The goal is to calculate the predicted the rating $R_{a,j}$ for the active user u_a for the item i_j. To do that first we need to form the neighborhood of users similar to the user u_a. Following the assumption that similarity between users depends on how much they share in terms of items liked in the past, the algorithm calculates similarity between user pairs to form user-user matrix. Then users with the highest similarity to u_a are called the neighborhood. Size of the neighborhood can be limited in two ways [Lekakos and Caravelas, 2006]. First, algorithm can decide to select only the top N most similar users, this ensures that our neighborhood is always big enough but there are no guarantees about the neighbors being similar enough. The other way is to set the similarity threshold and only select users that satisfy the chosen level, this makes sure that we have only the right set of the neighbors but then we may end up having too few of them. In the given example the closest neighbors happen to be users u_2, u_{m-2} and u_{m-1}.

Once the neighborhood is formed then the next step is to combine their ratings in order to predict ratings for the user u_a. As a result we then get a prediction for item i_j or top N recommendation for the user u_a.

In every recommendation system the main requirements are the quality of recommendation and scalability. The latter is especially important in systems with millions of users and items (Amazon.com, Last.fm, YouTube, etc). In the user-based collaborative filtering the neighborhood formation step - the user-user similarity matrix U - is calculated on-line due to the dynamic nature of the users. While not being a major issue in small databases, it really becomes a problem when we start talking about millions of users and items. The problem is that computational cost raises exponentially with every user and item. In order to address the scalability, another technique was introduced - item-based collaborative filtering [Sarwar et al., 2001, Linden et al., 2003].

Item-based collaborative filtering is very similar in terms of the structure of the process and techniques used, but it starts with the different assumption *the user will be interested in items that are similar to the ones he liked in the past*. Just like the user-based filtering it has two steps - neighborhood formation and ratings prediction. The only difference is that the neighborhood is not the users but the items. So instead of searching for the similar users, it searches for the similar items. One fundamental difference between user-based and item-based collaborative filtering is that in the case of user-based filtering the similarity is computed along the rows of the matrix (users), whereas in the item-based approach the similarity is computed along the columns (items) [Sarwar et al., 2001]. The main advantage of such change is that item-item similarities can be calculated offline due to their more static nature, whereas user-user similarities need to be calculated on-line thus presenting major scala-

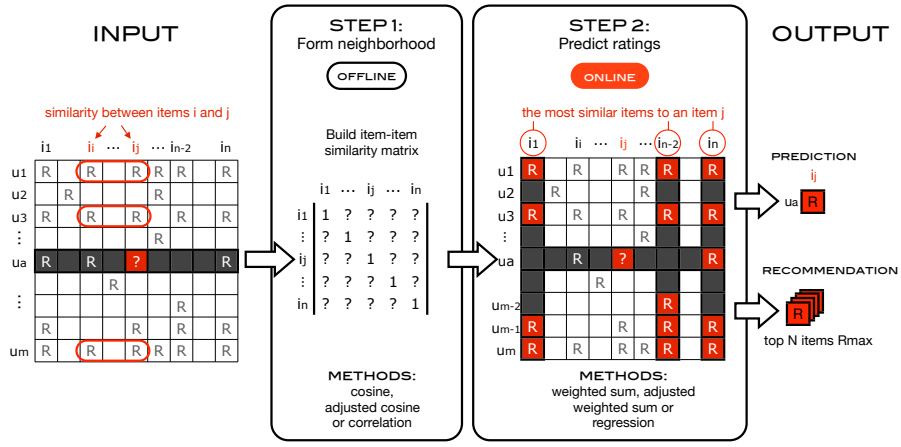


Figure 2.4: Item-based collaborative filtering precess.

bility issues. Therefore item-based collaborative filtering tends to produce much faster recommendations and is better suited for the cases where scalability is of an issue [Linden et al., 2003]. The algorithm's online component - looking up similar items for the user's purchases and ratings - scales independently of the catalog size or the total number of customers; it is dependent only on how many titles the user has purchased or rated [Linden et al., 2003]. Thus, the algorithm is fast even for extremely large data sets. Because the algorithm recommends highly correlated similar items, recommendation quality is reported to be better than the user-based method [Sarwar et al., 2001]. Unlike traditional collaborative filtering, the algorithm also performs well with limited user data, producing high-quality recommendations based on as few as two or three items. On top of that item-based technique has proven to outperform user-based alternatives in the domains where items are complex and multidimensional, for example music or movies [Huang et al., 2007]. The actual process goes in the following way (Figure 2.3). Item-based collaborative filtering starts with the same ratings matrix as the previously depicted user-based collaborative filtering. The only difference is that instead of calculating similarities between users, it calculates similarities between items resulting in the formation of the item-item similarity matrix. To calculate similarity between items i_j and i_i it first isolates all users who rated them both, then using standard similarity calculations it produces the number representing similarity between selected items. So far the process happens offline. The online parts starts by selecting the the neighborhood for the reference item i_j from the item-item matrix. Selected neighborhood's items ratings are combined using the same standard methods as in the user-based example. As an end result prediction or recommendations are generated.

2.3.2 Content-based filtering

While collaborative filtering derives all the knowledge from the interaction matrix, made by letting people rate items, it ignores the innate features an qualities of the items in question. Second big group of recommendation techniques is taking exactly the opposite point of view. This group is called content-based, and as the name implies it focuses on the content itself. Content-based approach starts with the same assumption as the item-based collaborative filtering - *the user will be interested in items that are similar to the ones user has liked in the past*. The main difference between them is in how the similar items are defined - the ones sharing the user base (item-based collaborative filtering) or the ones sharing features (content-based filtering).

In order to produce content-based recommendations, items have to be described by some features. To use item features as the main criteria in the recommendation process is an idea that comes from the fields of information retrieval [Baeza-Yates and Ribeiro-Neto, 1999, Salton, 1989] and information filtering [Belkin and Croft, 1992]. The features become of primary importance, the more descriptive they are, the more knowledge we can gain about the items and consequently the easier it is then to recommend items that are in some ways similar to the reference item, thus creating most value to the user [Lekakos and Caravelas, 2006].

Just like the collaborative filtering, content-based techniques can be memory-based or model-based. Memory-based techniques are usually based on the information retrieval methods, while model-based involve some kind of machine learning. In general content-based recommendations deals with two major problems: how to get the most descriptive features from content and then how to interpret them and use in the search for similar items (Figure 2.5). If we try to draw further parallels between information retrieval and content-based recommendations we could look at user profiles (or their usage history) as a query and then the problem is to retrieve the most relevant items to match the given query. In content-based recommendations as an input we still have item space I, it is just that items are no longer identified only by id but also have a set of features f associated with them. For example, in a movie recommendation application, where I is a collection of movies, each movie can be represented not only by its id, but also by its title, genre, director, year of release, leading actors, etc. [Adomavicius and Tuzhilin, 2005b]. Just like in the item-based collaborative filtering, content-based filtering is building an item-item matrix to identify similar items, but in this case it is the feature that are the main criteria for that. A content-based recommender learns a profile of the user's interests based on the features present in objects the user has rated. Schafer, Konstan and Riedl call this *item-to-item* correlation which is calculated by building the

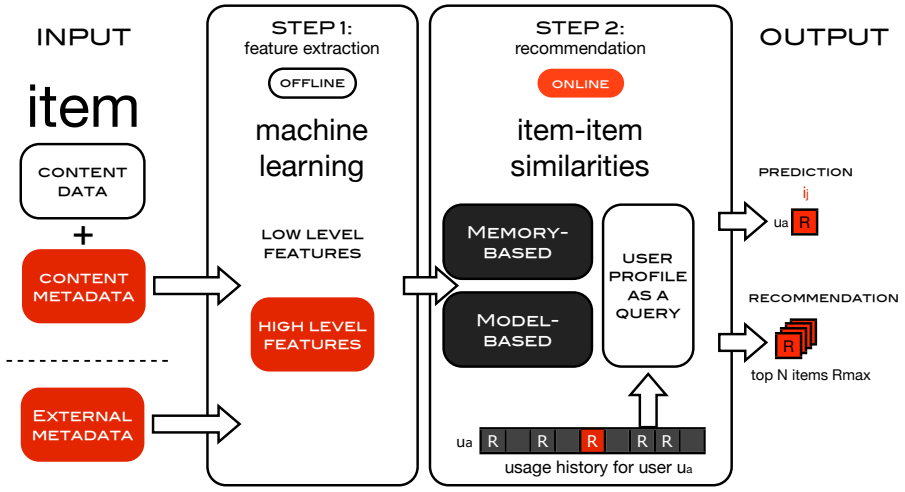


Figure 2.5: Content-based recommendation process

item-feature matrix in the background, from which we can then see all partial similarities that items share through sharing some of the features (Figure 2.6).

In order to build an item-feature matrix we naturally need to know what the features are. The features in the most general sense is all the information that we can possibly associate with the item. In many cases every item already comes with some extra information attached to it - the metadata. Metadata related features (which are discussed in the next chapter) can be anything from technical parameters of the media item, and all the way to manually added keywords and descriptions. If present, such information can be easily extracted because it is already in a machine readable format (serialized in XML). If such information is not present then features have to be extracted from the data itself. The complexity of this process depends on the complexity of the data - it is relatively easy for text, and very hard for complex media like audio and video. Since memory-based content recommendations build on the research done in the field of information retrieval, such systems are used mostly to recommend text-based items where the content is usually described with keywords [Adomavicius and Tuzhilin, 2005b]. Good examples of a content-based recommendation systems are the NewsWeeder [Lang, 1995], Fab system [Balabanovic and Shoham, 1997] or the Syskill & Webert system [Pazzani, 1997]. When it comes to more complex media, practical implementations are much more scarce due to the complexity of the problem, one of the few examples could be Blobworld [Carson et al., 2002]. Content-based retrieval from multimedia databases is an important key application

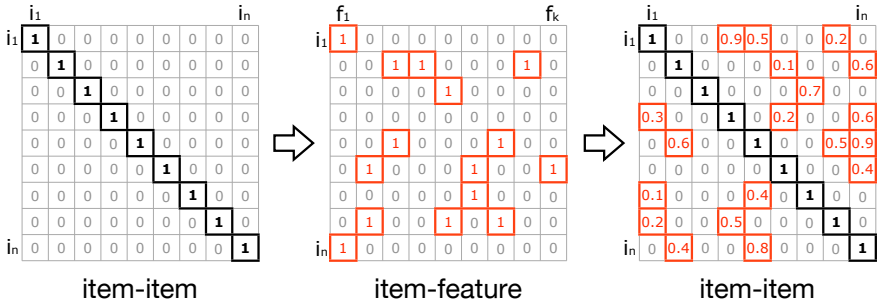


Figure 2.6: Item-item similarities inferred through the features.

[Gevers and Smeulders, 2004, Sebe et al., 2003], but despite the huge demand content-based search and recommendations are still very limited. The recent study of web search engines highlights that only 5 out of the 102 Web search engines support this feature [Tjondronegoro and Spink, 2008]. Even when content-based search is supported, users can only use low and mid level features. Research fields of content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) are main contributors to this area. The status at the moment is that automatic feature extraction is still very hard to achieve and due to the semantic gap can not be taken as a reliable source of information for complex media recommendations.

The good thing is that audio and video media is almost always annotated manually and already comes with a set of features eliminating the need to extract them from the raw data itself. Since all the media annotation standards are text based, we can apply standard information retrieval methods that we created to deal with the textual information. In text based information retrieval the item (text document) is viewed as a bag of features (words). Then we can to a certain degree infer what a document is like by looking at the words. The words are not equally descriptive and some of them just add noise rather adding information. The key idea is to add weights to every words and by doing that to separate the important words from noise.

One of the best-known measures for specifying word weights in information retrieval is the Term Frequency-Inverse Document Frequency (TF-IDF) [Salton, 1989]. Essentially, TF-IDF works by determining the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire document corpus [Balabanovic, 1997, Pazzani, 1997]. Assume that N is the total number of documents that can be recommended to users and that keyword k_j appears in n_i of them. Moreover, assume that $f_{i,j}$ is the number of

times keyword k_i appears in document d_j . Then, $TF_{i,j}$, the term frequency (or normalized frequency) of keyword k_i in document d_j, is defined as:

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}}$$

where the maximum is computed over the frequencies $f_{z,j}$ of all keywords k_z that appear in the document d_j . However, keywords that appear in many documents are not useful in distinguishing between a relevant document and a nonrelevant one. Therefore, the measure of inverse document frequency IDF_i is often used in combination with simple term frequency TFi_j. The inverse document frequency for keyword k_i is usually defined as:

$$IDF_i = \log \frac{N}{n_i}$$

Then, the $TF - IDF$ weight for keyword k_i in document d_j is defined as:

$$w_{i,j} = TF_{i,j} \times IDF_i$$

Another memory-based information retrieval technique comes from the field of Information Theory [Shannon, 1948]. The main idea still is to cluster or classify items based of their features. Like in TF-IDF, the the features (words) are treated differently based on their importance or informativeness. In this case the informativeness is calculated using the Information theory methods, namely Information Gain. The higher the information gain, the more significant the words in classifying documents. Information gain is equivalent to the mutual information that shows the expected reduction in entropy caused by partitioning the items i according to the feature f_p:

$$InfoGain(f_p) = H(i) - H(i|f_p)$$

$H(i)$ is the Shannon's entropy and can be calculated from the following formula:

$$H(i) = -\sum_j P(i_j) \log_2 P(i_j)$$

Shannon used the term entropy to quantify the information. It's meaning can be interpreted in many ways – it shows the amount of information, a level of disorder in the system, a level of the unpredictability of the next value that comes out of the information source. Formally the entropy is understood as the average amount of information that the observer has gained after receiving a realized outcome x of the random variable X. The base of the logarithm determines the unit of information. Since it is convenient to quantify information in bits, Shannon used logarithm base 2. It can be seen from the formula that the amount

of information in the information source depends on the sum of the individual probabilities of the symbols emitted from the information source. When the system takes only 1 value or symbol, then the entropy is 0, because we do not get any information finding out what is certain to happen. According to Shannon's theory, the amount of information is greater for a source that has a high level of uncertainty (is less predictable).

$H(i|f_p)$ is called the conditional entropy. In this context it represents the entropy of an item i given that it has a feature f_p:

$$H(i|f_p) = -\sum_j P(i_j|f_p) \log P(i_j|f_p)$$

The more informative the feature is, the more entropy of the item will decrease when we are told that the item contains the the feature f_p. To put it the other way, this technique focuses on selecting features which add most information (have the biggest Information Gain) and then using them for classification of items.

Both TF-IDF and Information Gain approaches are relatively simple but efficient when we are dealing with text, since in many cases it can be directly interpreted by machines. When we have to deal with more complex media it becomes extremely hard due to the semantic gap that such complex media presents. Then we need the machines to be able to understand and interpret media before making recommendations. The process of automatic annotation by mapping low-level features into high-level semantic concepts is generally difficult as it needs machine learning and interpretations [Tjondronegoro and Spink, 2008].

But before model-based recommendation methods are discussed, it is important to get back to one aspect of the recommendation process that has not been discussed yet - the exact methods used to calculate similarity between users or items (step 1), and how the rating are combined to calculate prediction or recommendations (step 2). The following section introduces the most popular methods to perform those operations. The methods apply not only to the memory-based filtering but to any kind of recommendations where we need to calculate similarity between two objects and to combine multiple rating values into one, therefore the methods introduced here apply in all the other recommendation approaches as well.

2.4 Calculating similarity and prediction

As previously stated, memory-based recommendations process can be divided into two individual steps: 1) find neighborhood and 2) calculate predictions and recommendations. This section explains how the actual calculation are being carried out.

Neighborhood building is the process of finding users or items that are similar in one way or the other. Ability to calculate similarity is one of most essential parts of the whole recommendation process. It is usually the case that recommended items are aimed to be similar to what user has indicated that she likes. Similarity can be understood in a number of ways, for example, in collaborative filtering we are talking about similar users because they share elements of their usage history, or similar items because they have a record being consumed together by the same people. In content based approach similarity refers to either sharing features (direct overlaps) or through containing similar features (partial overlaps).

There are many of ways to calculate similarity. The most popular methods implemented in the actual recommendation systems are either vector based or using Pearson-r correlation [Sarwar et al., 2001].

In the vector based method two items are thought of as two vectors in the multidimensional space. In case of recommendation systems we want to be able to compare users or items, therefore users and items are expressed as vectors: user is represented as a vector in the item space $\vec{u_a} = (i_1, i_2, .., i_m)$ while the item is represented as a vector in the user space $\vec{i_j} = (u_1, u_2, .., u_n)$. In case of content-based filtering every item is normally expressed as a bag of features. Instead of that we can imagine item being a vector in the feature space $\vec{i_j} = (f_1, f_2, .., f_k)$. Once we have our variables expressed as vectors then the similarity between them is measured by computing the cosine of the angle between these two vectors [Sarwar et al., 2001, Breese et al., 1998]. Vector similarity can be calculated from the following formula, where "$\cdot$" denotes the dot-product of the two vectors. This is known as *cosine similarity*.

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\| * \|\vec{j}\|} = \frac{\sum_{u\in U} R_{u,i}R_{u,j}}{\sqrt{\sum_{u\in U} R_{u,i}^2}\sqrt{\sum_{u\in U} R_{u,j}^2}}$$

Formula above shows cosine similarity between two items i and j, but it can similarly be applied to calculate similarity between two users or two features, as long as they are expressed as vectors. U is the set of users who have rated both items i and j, $R_{u,i}$ and $R_{u,j}$ are ratings given by the user u to items i and j respectively.

Computing similarity using cosine similarity formula in item-based case has one important drawback - the differences between different users in their rating scale are not taken into account. This is because when we search for similar items we take rating from different user pairs. This problem has been addressed by subtracting the corresponding user average from each co-rated pair [Sarwar et al., 2001]. The adjusted cosine similarity can be found from the following formula, where $\bar{R_u}$ is the average of the uth user's ratings.

$$sim(i,j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R_u})(R_{u,j} - \bar{R_u})}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R_u})^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R_u})^2}}$$

$$sim(a,m) = \frac{\sum_{i \in I} (R_{a,i} - \bar{R_a})(R_{m,i} - \bar{R_m})}{\sqrt{\sum_{i \in I} (R_{a,i} - \bar{R_a})^2} \sqrt{\sum_{i \in I} (R_{m,i} - \bar{R_m})^2}}$$

Another very popular approach is called *correlation-based similarity*. In this case, similarity between items i and j is computed using Pearson-r correlation [Sarwar et al., 2001, Shardanand and Maes, 1995, Resnick et al., 1994]. Where $\bar{R}_i$ and $\bar{R}_j$ are the average ratings for items i and j.

$$sim(i,j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_i)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_j)^2}}$$

Each one of those three popular methods produces numerical value that expresses the similarity between two items i and j. This number can then be used as a cell entry in the item-item or user-user matrix from which we select our neighborhood. Once the neighborhood is formed we need to combine their ratings in order to estimate the ratings for the user u_a. Two most popular techniques to do just that are weighted sum and regression model.

As the name implies, weighted sum method computes the sum of the ratings given by putting a weight next to each rating depending of how much two users or items are similar. In item-based approach similarity $sim(i,j)$ between items i and j is used as the weight [Sarwar et al., 2001], whereas in user-based collaborative filtering similarity $sim(u,u')$ between users u and u' is used instead [Adomavicius and Kwon, 2007]. If we take user-based approach as an example then there are two ways to apply weighted sum method:

normal weighted sum method

$$R_{u,i} = z \sum_{u' \in U} sim(u, u') \cdot R_{u',i}$$

and adjusted weighted sum method

$$R_{u,i} = \bar{R}_u + z \sum_{u' \in U} sim(u, u') \cdot (R_{u',i} - \bar{R}'_u)$$

Where $R_{u,i}$ the rating that user u would give to an item i, $\sim (u, u')$ is the similarity between two users u and u', $\bar{R}_u$ and $\bar{R}'_u$ are average rating of the two users.

Both methods weight the value of the rating $R(u', i)$ by the similarity of one user u to the other user u' - the more similar the two users are, the more weight $R(u', i)$ will have in computing the value of $R(u, i)$ [Adomavicius and Kwon, 2007]. Furthermore, multiplier z serves as a normalizing factor and is set to represent how big the similarity actually is:

$$z = \frac{1}{\sum_{u' \in U} |sim(u, u')|}$$

Another way to combine ratings is called regression. In practice, the similarities computed using cosine or correlation measures can be misleading in the sense that two rating vectors may be distant (in Euclidian sense) yet may have very high similarity. In that case using the raw ratings of the so-called similar item or item may result in poor prediction. The basic idea is to use the same formula as the weighted sum technique, but instead of using the similar ratings values use their approximated values based on a linear regression model.

Notice that the end result coming out of this step is the estimation of a rating $R_{u,i}$ user u would give to unseen item i. The estimated rating can be presented to the user as a prediction, or N items having the highest ratings can be combined to form a recommendation.

2.5 Model-based recommendations

In contrast to memory-based methods, model-based algorithms [Billsus and Pazzani, 1998, Breese et al., 1998, Getoor and Sahami, 1999, Goldberg et al., 2001, Hofmann, 2003, Ungar and Foster, 1998, Sarwar et al., 2000a, Mobasher et al., 2000] use the collection of ratings to learn a model, which is then used to make rating predictions

and recommendations. Algorithms in this category take probabilistic approach and differ from memory-based approaches in that they calculate utility predictions based not on a heuristic formula, such as a cosine similarity measure, but rather are based on a model learned from the underlying data using statistical learning and machine learning techniques, most of which are based on Bayesian principles.

Memory-based approach builds on the classical definition of probability, which comes from Pierre Simon Laplace from more than 200 years ago. As stated in his "Theorie analytique des probabilites": "The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible".

Thomas Bayes introduced another way to look at the probabilities by interpreting probability as a measure of a state of knowledge. The main difference between the classical frequentist and Bayesian approaches is that in the later the prior information is included in a calculation of an a posteriori probability while the non-Bayesian methods assume that one knew nothing of the thing being sampled prior to the sampling. Bayes' theorem relates the conditional and marginal probabilities of stochastic events A and B and looks like this:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $P(A)$ is the prior probability or marginal probability of A, $P(A|B)$ is the conditional probability of A, given B (it is also called the posterior probability because it is derived from or depends upon the specified value of B), $P(B|A)$ is the conditional probability of B given A and $P(B)$ is the prior or marginal probability of B which acts as a normalizing constant. Bayesian principles are used in a number of different ways in machine learning. The most popular techniques are the Bayesian Clustering, Bayesian Networks and Maximum Entropy.

Bayesian clustering is used for both collaborative and content-based recommendations. In the collaborative filtering approach, the clustering model [Pazzani, 1997, Osinski and Weiss, 2005] is used to cluster like-minded users into classes. Given user's class membership, user's ratings are assumed to be independent (it means that the model structure is that of a naive Bayesian network). The number of classes and the parameters of the model are learned from the data [Pennock and Horvitz, 2000]. Clustering model treats collaborative filtering as a classification problem [Breese et al., 1998, Basu et al., 1998, Ungar and Foster, 1998, Mobasher et al., 2000] and concentrates on estimating the probability that a particular user is in a particular class, and from there computes the conditional probabilities for ratings. This process has two steps: first the algorithm assigns the user to the segment containing the most similar customers, then it

uses the purchases and ratings of the customers in the segment to generate recommendations[Linden et al., 2003, Ungar and Foster, 1998].

Some clustering techniques represent each user with partial participation in several clusters [Sarwar et al., 2000b, Ungar and Foster, 1998]. The prediction is then an average across the clusters, weighted by the degree of participation. Clustering techniques usually produce less-personal recommendations than other methods, and in some cases, the clusters have worse accuracy than the nearest neighbor algorithms [Breese et al., 1998]. Once the clustering is complete, however, performance can be very good since the group that needs to be analyzed is much smaller. Clustering techniques can also be applied as a "first step" for shrinking the candidate set in a nearest neighbor algorithm or for distributing nearest neighbor computation across several recommender engines. While dividing the population into clusters may hurt the accuracy or recommendations to users near the fringes of their assigned cluster, pre-clustering may be a worthwhile trade-off between accuracy and throughput.

In the content-based filtering the Bayesian classifier is used for the same purpose as in the collaborative recommendation, the difference is that the classification criteria is now the features of the content (usually expressed as keywords). To classify content, it simply assumes that an item's features are independent of each other [Ha, 2006].

Because optimal clustering over large data sets is impractical, most applications use various forms of greedy cluster generation. These algorithms typically start with an initial set of segments, which often contain one randomly selected customer each. They then repeatedly match customers to the existing segments, usually with some provision for creating new or merging existing segments [Bradley et al., 1998]. For very large data sets - especially those with high dimensionality - sampling or dimensionality reduction is also necessary. Clustering models have much better online scalability and performance than memory-based collaborative filtering [Breese et al., 1998] because they compare the user to a controlled number of segments rather than the entire customer base and the complex and expensive clustering computation is run offline.

Another machine learning technique used in recommender systems is called Bayesian Networks. A Bayesian Network is a probabilistic graphical model that represents a set of variables and their probabilistic independencies. Variables in the network are the items and their values are the allowable ratings. Both the structure of the network, which encodes the dependencies between titles, and the conditional probabilities are learned from the data - the model is build. Bayesian networks create a model based on a training set with a decision tree at each node and edges representing user information. The model can be built offline over a matter of hours or days. The resulting model is very small, very fast,

and essentially as accurate as nearest neighbor methods [Breese et al., 1998]. Bayesian networks may prove practical for environments in which knowledge of user preferences changes slowly with respect to the time needed to build the model but are not suitable for environments where user preferences must be updated rapidly or frequently.

Maximum entropy (also called maxent) model-based technique focuses on selecting content features with maximum entropy [Garden and Dudek, 2005] and has been implemented in several recommendation system [Jin et al., 2005, Pavlov et al., 2004]. It has been widely used in Natural Language Processing, Information Retrieval, Text Mining and other areas. Since entropy refers to the level of informativeness and surprisal, maximum entropy means that the system is least predictable, and that we are the least informed about the possible outcome. In his famous 1957 paper [Jaynes, 1957], E. T. Jaynes wrote: "Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum entropy estimate. It is least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information. That means that, when characterizing some unknown events with such statistical model, we should always choose the one that has Maximum Entropy.

2.6 Hybrid methods

In order to exploit the advantages of all the recommendation methods presented up until now, several hybrid approaches have been proposed, in their vast majority concerning combinations of content-based and collaborative filtering [Balabanovic, 1997, Burke, 2002] or extension the two methods by demographics-based predictions [Pazzani, 1999], while few of them utilize knowledge-based techniques where domain functional knowledge is exploited [Burke, 2002]. Researchers have also started to combine both memory-based and model-based approaches. Memory based techniques generate predictions but do not explain why they are such, model-based approaches on the other hand generate explicit assumptions and have a meaningful probabilistic interpretation as to why certain prediction was made [Pennock and Horvitz, 2000].

Burke [Burke, 2002] classifies hybridization techniques into seven classes (Figure 2.7). The most popular way to improve performance and quality of the recommender is to utilize both collaborative and content-based methods in order to avoid certain limitations that each method presents [Ungar and Foster, 1998, Schein et al., 2002, Pazzani, 1999, Basu et al., 1998, Balabanovic and Shoham, 1997]. Therefore the first three hybrid approaches deal exactly with a combination of

content-based and collaborative methods. The first one is called *weighted* where each of the recommendation approaches makes predictions which are then combined into a single prediction; *switching* where one of the recommendation techniques is selected to make the prediction when certain criteria are met; *mixed* in which predictions from each of the recommendation techniques are presented to the user.

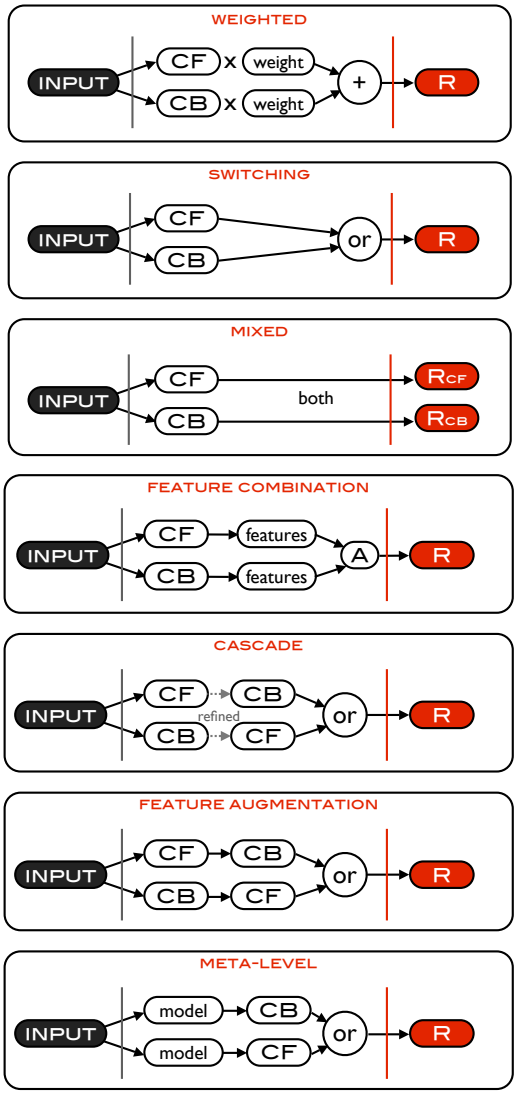


Figure 2.7: *Weighted*, *switching*, *mixed*, *feature combination*, *cascade*, *feature augmentation* and *meta-level* hybrid approaches.

Switching, mixed, and weighted hybrids are differentiated from the remaining techniques in Burke's taxonomy by the fact that each of the individual (base) recommendation methods produce independently from each other a prediction which is then presented to the user either as a single combined predic-

tion (switching, weighted) or as two independent predictions (mixed). Switching hybrids in particular, are low-complexity hybridization methods based on the examination of the conditions that affect the performance of the base algorithms each time a prediction is requested. [Lekakos and Caravelas, 2006]

The other four approaches are more complex and incorporate different model-based features. They are the following: *feature combination* where a single prediction algorithm is provided with features from different recommendation techniques; *cascade* where the output from one recommendation technique is refined by another; *feature augmentation* where the output from one recommendation technique is fed to another, and *meta-level* in which the entire model produced by one recommendation technique is utilized by another.

There are several important extensions recently introduced in the area of recommendation systems. One of them is knowledge-based approach to recommendation. The knowledge-based approach is implemented by collecting information (through dialogue) concerning product features and related importance value and subsequently exploiting domain knowledge to match products to the user needs. Although the knowledge-based process is resource intensive requiring human knowledge-engineering effort, it can provide recommendations even in the complete absence of user ratings. Usually the knowledge-based approach is applied in the following way. The system monitors the number of users with known interest profiles and the number of rated items in the database. If either of the two variables is below a fixed threshold, then the knowledge-based recommendation is presented to the user, otherwise the collaborative approach is applied. [Lekakos and Caravelas, 2006]

Another improvement comes from the understanding that items which are being recommended almost always are complex enough to be rated separately for each of their key features. The vast majority of current recommender systems use a single criterion, such as a single numerical rating, to represent an item's utility to a user in a 2D *Users-Items* space. Single-criterion rating systems have proved successful in several applications, but many industries have begun employing multicriteria systems. This move indicates that multicriteria data provides value to online content providers and consumers as a component in personalization applications. [Adomavicius and Kwon, 2007]

Overall rating shows *how much* user likes certain item. Multicriteria approach shows *why* user likes the item. In user-based collaborative filtering this enables to estimate more accurate similarity between two users and results in a better neighborhood which in turn improves recommendation quality while in the item-based collaborative filtering and content-based filtering, this enables to find items that are similar not necessarily overall but according to some criteria, then this criteria can be reflected in user profiles and it makes recommendations

more accurate and novel.

The potential of the multicriteria ratings have been shown through a number of different implementations. Music recommendation system *RACOFI* lets users rate contemporary Canadian music in the five dimensions of impression, lyrics, music, originality, and production [Anderson et al., 2003]. In the resource discovery *CoFIND* [Dronf et al., 2000] system the user provides feedback on the qualities of the item and qualities are suggested if they are used repeatedly. In the *Entree* system [Burke, 2002], Burke created a recommender system for restaurants in which user feedback is not specified as a numeric overall rating, but in which users specify a semantic rating in which he or she feels the current item is lacking. In the *Recommendz* movie recommendation system, users are allowed to specify explicitly which aspects of the movie they liked most [Garden and Dudek, 2005, Garden and Dudek, 2006]. Multicriteria approach is also applied in a publication search system *CiteSeer*[Pavlov et al., 2004] by recommending other publications to the user based on different parameters: text similarity, citation similarity, etc.

2.7 Problems of recommendation systems

The huge research interest in the field of recommendation systems remains because there are still many unsolved problems. Despite being the most successful recommendation techniques to date collaborative filtering still suffers from numerous issues:. The main ones are:

- new user
- new item
- sparsity
- attacks and remedies

The main problem comes from the fact that all the collaborative filtering recommendations are made from the interaction matrix alone without any explicit knowledge about users and items. That means that a new user who just started using the system will not get any recommendations because her usage history is empty and thus the system does not have any information about the user to generate recommendations. This is known as the new user problem (also called cold start) [Schein et al., 2002, Smyth and Cotter, 1999, Sollenborn and Funk, 2002]. Similarly on the content side all we know about

the item is which users have consumed it and their ratings, if they are available. This leads to new items being out of reach up until sufficient number of people somehow find it and rate it. It is known as a first rater problem [Balabanovic and Shoham, 1997, Konstan et al., 1997, Pazzani, 1999].

Another big problem is the sparsity of the ratings matrix. The sparsity directly influences the neighborhood formation and thus is a major scalability bottleneck. This problem has been addressed in numerous publications [Good et al., 1999, Billsus and Pazzani, 1998, Sarwar et al., 2000b] and a number of solutions have been proposed. *Generative models* use latent class variables to explain the patterns of interactions between consumers and products [Hofmann, 2004], *spreading activation algorithm* efficiently explores the connectedness of a consumer-product pair in the consumer-product graph [Huang et al., 2007]. The one of the most promising techniques used for dimensionality reduction of the ratings matrix is Singular Value Decomposition (SVD). SVD algorithm condenses the original interaction matrix and allows to generate recommendations based on the condensed, less-sparse matrix to alleviate the sparsity problem. SVD decomposes the interaction matrix A of size $m \times n$ into three matrixes $U \times Z \times V$, where U and V are two orthogonal matrices of size $m \times r$ and $r \times n$ respectively, and r is the rank of the matrix A [1]. Z is a diagonal matrix of size $r \times r$ having all singular values of matrix A as its diagonal entries. The algorithm then reduces Z by retaining only the k largest singular values, to obtain Z_k. It reduces U and V accordingly to obtain U_k and V_k. So, $U_k \times Z_k \times V_k$ provides the best lower-rank approximation of the original interaction matrix A that preserves the primary data patterns existing in the data after the noises are removed. Consumer or item similarities can then be derived from the compact representation.

Recently another problem has appeared in the area of collaborative filtering - attacks and remedies [Hurley et al., 2007, Mobasher et al., 2007]. Since the only input into collaborative recommendations is the ratings matrix, the quality of the recommendation depend on the accuracy of the data in the matrix. Matrix itself is very accessible to anybody through their own usage history, every user makes contribution and has impact on recommendation process. Attacks and remedies occur when users generate false profiles filled with false ratings and that in return may cause recommendations to be less accurate.

Content-based techniques do not suffer from new item problem because they rely on the features and not on the ratings matrix, therefore we do not need to wait till the critical mass of ratings accumulate. Nevertheless, content-based recommendations have their own problems:

[1] An example of the SVD process is illustrated in the Figure 6.3

- new user
- limited content
- overspecialization

New user problem remains even though it is not as big as in the collaborative filtering. It arises from the fact that the recommendation system first needs to learn new user preferences, to understand user's needs. That means that the user must to consume (or rate) a number of items before the features of those rated items can form an understandable picture of what the user needs are like.

Limited content problem is directly related with the features that the recommendation system uses. We can say that content-based techniques are limited by the features that are associated with the items that these systems recommend. If an item has too few features associated with it, then it is very hard to really understand what this item is all about and to recommend it. In order to have a sufficient set of features, the content must be either in a form that can be parsed automatically by a computer (e.g., text) or the features should be assigned to items manually [Adomavicius and Tuzhilin, 2005b]. While information retrieval techniques work well in extracting features from text documents, some other domains have an inherent problem with automatic feature extraction. For example, automatic feature extraction methods are much harder to apply to multimedia data, e.g., graphical images, audio streams, and video streams. Moreover, it is often not practical to assign attributes by hand due to limitations of resources [Shardanand and Maes, 1995]. Another problem with limited content analysis is that, if two different items are represented by the same set of features, they are indistinguishable. Therefore, since text-based documents are usually represented by their most important keywords, content-based systems cannot distinguish between a well-written article and a badly written one, if they happen to use the same terms [Adomavicius and Tuzhilin, 2005b][Shahabi and Chen, 2003] [2].

Every time researchers compare collaborative filtering with the content-based methods, they usually highlight overspecialization as the main problem of content-based approach and is completely absent in collaborative filtering. Overspecialization means the lack of diversity [Shahabi and Chen, 2003][Adomavicius and Tuzhilin, 2005b] in terms of content items being presented to the users. The usual examples that are given to illustrate this problem are the ones where user is limited to the movies of the same genre or featuring the same actors that user liked in the past or music by the same bands. Since novelty is one of the main requirements

[2]This problem is usually solved by using the "network effects" as a mechanism to let people to separate media items in terms of perceived quality. Such effects are the very core of collaborative filtering process

for the recommendation process, overspecialization is a serious issue. This problem has it's roots in the features themselves. How can the system ensure the diversity of the items, when the feature space connected with the items is not diverse itself? Therefore the problem can be solved by improving the quality and quantity of the content descriptions (the features). Another way to address this issues is to think of other novel approaches of how to extract diverse features from the content descriptions. One such approach is presented in this thesis and focuses of extracting features that are based on emotions from the widely available content metadata.

Even though it was mentioned before, scalability is an issue worth revisiting again. First let me remind that there are on-line and off-line environments. The main bottleneck naturally comes from the computational time spent in the on-line environment. Traditional collaborative filtering is calculating the neighborhood of the most similar users on-line, and that in return limits the scale to which such approach can be applied. Item-based collaborative filtering on the other hand computes the item-item matrix off-line and thus scales much better. Model-based approaches also scale much better compared with traditional filtering since the most computationally expensive process - building of the model - is performed off-line. Generally the content based methods scale much better than traditional collaborative filtering because, much like item-based techniques, the most time consuming part of the process is done offline. That part includes building an item-item matrix from which we see which items are similar to which ones. The problem though is how to get features that are descriptive enough. In domains such as movies, videos, or music, features extraction is hard to achieve due to the semantic gap [Balabanovic and Shoham, 1997].

2.8 Conclusions

To conclude, personalization system contains multiple parts (Figure 2.1). Media content, user profiles, contextual information and various business rules and service policies are the input into the system. The actual processing is performed by a recommendation engine that produces personalized output - a set of media items highly relevant for the given user in a given situation.

Traditional collaborative filtering is strong when it comes to producing highly diverse and novel recommendations. It draws it's strength from relying only on the ratings matrix and all the user-item interaction data that can be extracted from it. The main challenge is then to select the best possible neighborhood for a given task. Due to the dynamics of the user profiles, such calculation must be done online, and this is the main scalability drawback of traditional

collaborative filtering. Item-based collaborative filtering addresses scalability issue by focusing on item-item correlations, instead of user-user similarities. Item-based collaborative approach is used in such scalability sensitive projects as Amazon.com. Both collaborative approaches ignore the features of items and users. The positive thing about it that the methods are applicable in any domain since they do not require any specific knowledge about the items.

Depending how the actual recommendation calculation process is carried out, both collaborative filtering content-based filtering can be memory-based or model-based. Memory-based is the more simple traditional method, that takes the entire input data and then calculates recommendation, while model-based approaches use the input data to build a model first and then use the model to calculate recommendations. Model-based approaches use one of the machine learning methods, most of which build on bayesian theorem.

When it comes down to the actual calculations, recommender systems first need to be able to calculate similarity between either users or items in order to build the neighborhood. Secondly they need to combine multiple values into one recommendation. Popular techniques used for the similarity calculations are either cosine based or correlation based, whereas the most popular methods for combining multiple ratings are weighted sum, adjusted weighted sum and regression.

Another recommendation approach pays attention to the features instead of only the ratings. It is called content-based filtering and it looks at items as a bags of features. The complexity, and most of the problems, here are directly connected with the features, their quality and diversity. As it comes to getting the features, they can be either provided explicitly or can be extracted from the content. While it is relatively easy to extract features from text (in this case features are in the form of keywords) it is extremely hard to do so when media becomes as complex as audio or video. On the other hand we do not necessarily need to extract features from raw data, we can also extract them from various textual descriptions of the media that are available. Examples of such descriptions include synopsis for movies, lyrics for songs, or simply reviews that express how users and critics feel about the media item.

Collaborative and content-based approaches are usually combined into a hybrid system to utilize their strengths. This also means that in most cases we will have features of items present, thus improving the quality of the features improves the overall process of recommendation. The following chapter presents the media features themselves - the state of the art in describing media.

Chapter 3

Describing Media

Even though media personalization is heavily researched area and we already have many media recommendation systems in place, there is still plenty of room for improvement. The biggest challenge is to overcome the so-called *semantic gap*. Since the more complex the media - the bigger the gap, most problems occur when we try to personalize complex media like audio and video. Even though simple memory-based collaborative filtering systems are currently dominating the market, they can not cope with very complex tasks that audio-visual media presents. This is not because the collaborative algorithms are not efficient enough, it is because they are build to rely on very limited input - the interaction matrix - ignoring the features of the media itself. While this is one of their strengths (they are easy to implement and make the most out of limited information), it is also their biggest weakness since it sets very clear limitations as to how intelligent such methods can become.

In order to improve media personalization the most intuitive idea would be to improve the recommendation algorithms that produce personalized content. Judging from the state of the art in recommendation systems, it can be stated that there has not been any ground-braking improvements in the last few years. All the new contributions in the area seem to be based on the old ideas and basically try to squeeze the most out of what may have already gone pretty dry. It leads to a conclusion that the recommendation algorithms themselves are not the main bottleneck stopping us from getting to the next level of media

personalization. My hypothesis here is that we can make more significant improvements if we focus on providing the existing algorithms with better input rather than focusing on the algorithms themselves. Or to put it in another way, no improvements on the algorithm side will compensate the lack of knowledge about the media itself.

In the previous chapter the input into the media personalization process was divided into two classes: the interaction matrix and the features of media. Interaction matrix is the main input into collaborative-based algorithms while the features form the basis for content-based filtering. Most of the media recommendation systems try to employ one or the other kind of hybrid recommendation approaches combining aspects of both content-based and collaborative filtering. Therefore in most cases recommendation system has (or potentially may have) a content-based side already, but this is also the side that completely depends on the availability and quality of the content side metadata and can go only as far as the metadata allows it.

The purpose of this chapter is to explain the role of the media metadata and to present the state of the art in media metadata standards, assessing their potential for media personalization. The main focus still remains on audio and video media since those two media forms present most of the challenges.

3.1 Media metadata

In the previous chapter the term *features* was used to identify the inner qualities of a media item, here the features are called *media descriptions* or simply *metadata*. Metadata (sometimes also called metainformation) is not a trivial term since it can mean different things in different situations. Most popular definitions of metadata are quite abstract "information about information" or "data about data". To put in a more concrete way:

> *"Metadata is structured, encoded data that describe the characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities"*[Durrell, 1985].

As can be seen from the definition, metadata is not limited to media or digital goods. It can be (and is) applied for every information-bearing entity - literally that includes every single physical or digital object that carries some kind of information. In the case of media such objects are songs, movies, video clips,

news articles, pictures, etc. All the extra information about those objects can potentially become metadata.

Before we dive into the actual practices and explore *how* media is described, it is worthwhile to take a step back and to ask two questions:

- why do we describe media?
- what user needs media intends to satisfy?

3.1.1 Why do we describe media?

A quick and direct answer to this question could be to say that media descriptions enable search and enhance recommendation. But since search and recommendation are two completely different ways to interact with media they each have their own reasons to have media descriptions.

As it was stressed in the *Chapter 2*, the idea of recommendation is very much about the novelty of the items. Therefore in this case we describe things in order to give some understanding to a person about the item without her having to consume it first. This is the main reason why TV programs have synopsis, movies and music albums have previews and reviews, etc. Here the purpose is to inform and to promote new media item by revealing certain information about it, that will help the user to form an opinion whether or not she is interested in actually consuming the item. This is very essential for media that is made to be consumed by as many people as possible, for instance, all the commercial content. On top of that the abundance of the novel media items appearing all the time makes it impossible to consume everything, therefore people need filters and recommenders to guide them to what they suspect to be worth their time and attention (see chapter 1.1). And since it is much faster to read the description of a movie compared to sitting down and watching it, media descriptions help us to save a lot of our time. This of course is in the case we are the ones making a decision, which is not not so efficient in the first place given the amounts of media out there. If the decision is made by a machine then the availability of media descriptions will determine if we are limited to only traditional collaborative recommendations or if we can employ more advances hybrid and content-based methods.

Another scenario is when people are looking for something that they already know exists. This type of action is known as the search. The importance of describing media in this scenario is even greater than in the recommendation case since the search becomes nearly impossible if media does not have a single

feature that can be referred with in a query. The search as such is widely used not only in commercial context but also to navigate in private media collections as well. All our pictures, movies, songs and other documents need to be either put in a very organized folders (librarian approach) or must be described with some features in order to be easily retrievable upon request.

Therefore recommendation and search functionality requires for the media to be described in one way or the other.

3.1.2 What is media's intention?

This is the second big question that gives a better insight into media personalization requirements in term of metadata. This is a big philosophical and sociological question and one could easily write a thesis just on that topic alone. Therefore I do not attempt to answer this question fully in this thesis, instead I will jump straight to conclusions/assumptions based on simple reasoning.

When people consume media their desire is usually to either seek for information or entertainment, or in some cases both. Without going into any further classification of neither informational nor entertainment media I will just give a few examples of each. Information-driven media examples would be news or all kinds of educational programs. Entertainment-driven media is much more diverse and contains many different layers and levels - music, movies, TV shows, sitcoms, games, sports, etc. And then there many cases where the purpose of the media is somewhere in between information and entertainment where information is presented in a very engaging and gripping way rather than just being some evening-news-type facts.

Looking from media personalization perspective information-based media is easier to personalize (compared with the entertainment media) since it is more about the facts and objective information rather than people's taste which is based on subjective criteria. Information media is usually annotated with metadata that describes what media *is about* rather that *what it is like* or *how it feels*. Therefore usually it relies on keywords describing key topics, places, events or people who appear in the media. Such annotation is relatively easy and straightforward, but it is extremely time consuming and thus expensive if we do it manually. Lots of automatic methods were proposed to help with this task. They mostly focus on keyword extraction either from the raw media file, or from supplementary material that comes with it - for example subtitles. Most contributions here come from the areas of image, audio and video processing, voice and face recognition. Since information-based media is more objective and facts oriented, the links between such content tend to be more static, since they

depend less on the user and more on the content.

Entertainment media is much more complex for the machines to understand. Even though there might be some informational value in it, such media primarily aims to make people to experience certain emotions - love, anger, fear, suspense, etc. This is very much the case with movies, music, theatre plays, operas, popular magazines or computer games. Music is perhaps the most emotional media that we have. There are so many things that go under the surface in music and it is very hard to try to point out the exact reasons why people like certain song. Even when we come up with the list of reasons, those are usually based on the different emotions that a song triggers in us. As a proof that this is the case we can simply look how people annotate music in any of the online music social networks where people are allowed to assign certain words to songs. We see that people do use emotional terms like love, cool, calm, soft, dark, etc. [Corthaut et al., 2006, Geleijnse et al., 2007]. The hard part is to quantify and map out such emotional terms in some space that we use to evaluate the similarities between different songs [1]. In order to successfully model recommendation systems that take emotions as a recommendation criteria, one needs to some extent to understand how the human brain works. This is the area of cognitive sciences, methods of which are theoretically discussed in *Chapter 4* and practically applied in *Chapter 6* and *Chapter 7*.

3.2 How do we describe media?

Now that we know why we need to describe media and that it has informational and emotional ways to interact with users, it is time to see exactly what information is used as the metadata in media? There is more than one way to classify media metadata. In the context of this thesis it is sufficient to simply group it into three classes:

- editorial metadata
- technical metadata
- content metadata

Editorial metadata refers to the information that assists in the location of the media, intellectual property rights, related material, special content packages

[1] "similarity" is still one of the key concepts for recommender systems, just in this case we could represent items as vectors in a semantic emotional space and then apply traditional techniques to determine similarity and generate predictions (see section 2.4)

and similar issues concerning the storage and distribution of the media. Editorial metadata is more typical in a commercial context and is usually outsourced to companies that specialize in this area (for example Red Bee Media, UK). Due to the scope of this thesis, editorial metadata will not be discussed any further.

Technical metadata is the most common type of media metadata simply because without it the consumption itself would not be possible. It consists of information about the item itself (ignoring what information is being stored in the item) and therefore depends only on three things: the type of the media (text, audio, video, etc.), the characteristics of the distribution platform and all the physical devices involved in the creation and consumption of media. Basically, technical metadata is all the low level information used to instruct user's device how to access and play media. Such metadata is usually generated automatically either by the device that is used to create media (photo camera, audio recording equipment or simply a computer) or by the network (for example the Internet). Despite its overall importance, technical metadata does not make significant contributions to personalization of media because it describes the physical characteristics of the item itself without giving us any clue what information it contains inside.

Last but certainly not least is the content metadata. It contains all the information that we use to get the insight about what the media item actually contains. Without any doubt it is the most valuable type of metadata for the personalization process since it represents all the qualities that humans consider to be important for classification and retrieval of media. To put it in another way, content metadata represents the user needs on the content side, whereas technical metadata represents what is important for the machines. For example, if we take an audio file, the machine needs such information as the exact location of the file, it's sample rate, volume level, coding, etc. All this data is relevant in order to play the song, but completely meaningless for the user. The way the user sees the same audio file, is that this is, for example, a song by "The Birthday Massacre" called "Kill The Lights" belonging to a genre *Gothic* and also tagged with keywords *industrial*, *electronic* and *female vocalist*, and if the user is lucky, on top of that she will find a short review about what it sounds like. Content metadata does not depend on any parameters of distribution platform or device - author and genre information do not change whether the song will be broadcasted or put online, as well as it does not depend on what kind of device will be used to create or consume it. In other words, content metadata holds the information that humans usually use to to identify, classify and evaluate media. As can be seen from this simple example, content metadata can be further divided into two groups: the information that helps to identify or classify media item based on objective parameter (the band name, title of the song, year of production, etc), and the one which intends to classify or evaluate media based on someone's subjective opinion (genre, keywords, review, etc).

The "objective" content metadata is always very clear and is meant to be understood directly. We could imagine this as all the content related information that is available even before anyone actually sees the movie or listens to a song - the names of all the people connected with the production of the media, the title of the content, the language and the subtitle information, country of production, etc. First of all such information serves as a unique fingerprint identifying the media item. *Title* alone, in many cases, is not enough since there are many cases of different media items sharing the same title. While this is not as big problem in video content, in the area of music song title does not help much to identify a song, for example in a music portal Last.fm there are several hundred songs called "I Love You". But if we take more information and put it together, we get a unique identifier for very piece of media. Secondly, such metadata is used for rule-based filtering that is more useful when we want to build customized playlists instead of getting novel recommendations.

The second class of content-related metadata attempts to answer a question "What is the content like?". As can be expected, any answer to such question is likely to generate a response "Says who?", therefore such metadata is extremely subjective. The most popular examples of such metadata are: Genre, Keywords, Synopsis, Comments, etc. When it comes down to describing what something is like, one could either try to classify the object using a predefined categorization where every category has its unique meaning (for example, Genre), or one could try to describe an object using plain words (for instance, Keywords and Synopsis) and leave it up to the people to make up their own meaning out of them.

3.2.1 Genres

The word "genre" comes from the French (and originally Latin) word for "kind" or "class". Today genres are used very widely in the areas of audio and video content. We are used to such movie genres as *drama*, *comedy* and *science fiction*, or music genres as *pop*, *rock* or *classical music*. We even think that we understand what they all mean, and sometimes once asked what kind of movies or music we like we simply name a genre hoping that the other person will understand exactly what we mean [2]. The truth is that "genres are vague categories with no fixed boundaries" [Chandler, 2003]. How much information do we really get about media upon being told what genre it belongs? In most

[2] to describe ones taste by using the genre information is more typical to the more advanced (expert) users who know the domain well enough. Cognitively this is known as prototype model. Another approach is to use exemplar model, meaning that the user would rather list several examples of the movies or music bands that she likes instead of giving one overall description of her taste (for more on this topic see the chapter 4).

cases, not much, because genres are either too broad or too narrow. Genres like *drama* or *metal* are so broad in terms of variety of the members they contain that they do not give many clues about what to expect from the content. On the other side of the spectrum, genres like *sadcore* [3] are much more descriptive but are so obscure that very few people actually know it. According to the Wikipedia, there 1552 unique music genres [Wikipedia, 2008], while one of the leading audio-visual content description specification TV-Anytime has defined 922 genres only for TV content [ETSI, 2004b]. Apart of being so numerous genres also have tendency to change, merge or disappear while the new ones are appearing all the time [Chandler, 2003]. Overall problems with using genres to classify media can be grouped in four different types: extension (genres being too broad or too narrow); normativism (when people have preconceived ideas of what kind of criteria media needs to satisfy in order to belong to a certain genre); monolithic definitions (when we think that media item can belong to only one genre); biologism (when people believe that every genre evolves through a standardized life cycle) [Stam, 2000]. Another big issue with the genres comes from the fact they rely on having a ground truth telling what genres actually mean and how they relate, and yet there is no universal solution to that [Crowston and Kwasnik, 2003]. There is a number of specifications that define to set such ground truths, but they are all different therefore fragmenting the audience that uses genres and makes the interexchange if not impossible then at least very hard. One might even argue that there will never be a common ground truth since genres by definition are vague and too subjective. The best chances for a certain genre model to take over the stage is if it has a critical mass of users and becomes a de-facto standard. Despite all the problems with the genres, they are still perceived by many as one of the main media classification criteria. Chapter 5 explores the potential of genres to be used for media recommendations in a TV domain.

3.2.2 Keywords

Another kind of content metadata is the keywords. While media genre metadata is usually a complex taxonomy with strict hierarchy and relations between different genres, keywords do not have any other structure other than given by the language itself. They are just stand-alone words each attempting to describe a certain quality of the media item. Metadata considered as "genre" in one system can be used as a "keyword" in another. The main difference remains in the underlying structure defining how the meaning will be inferred and not in the actual word itself. Another big difference between these two content metadata types is that in order to assign genres one needs to have certain level of expertise

[3] "Sadcore" is an subgenre of alternative rock characterized by bleak lyrics, downbeat melodies and slower tempos

to be able to put things in the right places especially when the system allows an item to belong to only one genre, whereas keywords are much easier to annotate and usually have no limitations to how many can be present at a time. On the other hand when it comes to inferring meaning, the interpretation of genre terms is much more straight forward due to the structure that they follow (that does not mean that genres are more meaningful). Interpretation of the keywords can be extremely tricky since usually there are no formal rules telling the precise meaning of a certain keyword and its relationship with the other keywords. One could imagine genres as a representative of order - structured, standardized, but inflexible and requires expert knowledge, whereas keywords are more from the world of chaos - very loose, do not require any expert knowledge and are easy to use, but hard to interpret and unpredictable. This definition also highlights the reasons why keywords have become the most popular way to annotate media for common users. User generated keywords are often referred to as "tags" and form folksonomies, which are discussed in the section 3.7.

3.2.3 Free-text metadata

The last big class of content metadata is all sorts of textual descriptions where we use the common language to describe media and it ranges from a few sentences to few paragraphs or even pages. In the TV world such metadata is called *Synopsis*. In movies and music industry it is usually in a form of *Previews* and *Reviews*. Or in general it can go under the names like *Comments* or *Description*. This type of information is meant to be read by humans themselves, that is why it is described in sentences using everyday language, knowing that humans are able to understand and interpret such information seamlessly in their brain. Machines are not quite at the same level of comprehension as humans therefore it is very hard for computers to understand such information. Nevertheless free-text metadata fields hold a lot of potential, where the main problem still remains in the area of language comprehension. From the recommendation system perspective this requires an advances model-based techniques where a certain model is used to infer meaning from language. One technique that stands out among others in this case is the LSA (Latent Semantic Analysis). LSA is a technique normally used in a natural language processing, but it can be also taken as a theory for knowledge acquisition. In the *Chapter 6* two different cases are presented (TV programs and music songs) where LSA is used as a method to extract emotional patterns from the free-text metadata fields of the respective media.

3.2.4 Content metadata and the personalization

All the metadata discussed up until now can be looked at as an extra information serving as a description of the media and is not part of the content itself. It seems logical, because the metadata by definition is somewhat dis-attached from the content it describes. But there are several cases where it gets a bit more complicated than that. It turns out that sometimes a piece of information can be both metadata and content and the same time. Two typical examples of that would be a song lyrics and a movie script. One advantage here is that such metadata is always present and does not cost anything since it is already part of the content. In both cases such information can be easily found on the Internet, few notable examples include iMSDb, claiming to be the largest movie scripts resource on the internet (http://www.imsdb.com), and as for music lyrics there are hundreds of sites offering such information absolutely for free (www.lyrics.com, www.azlyrics.com, etc.). Another advantage is that such metadata gets as close to the content as possible, since it is part of it. And since just like the rest of the metadata it is in the form of text (rather than being raw audio-visual features) it can be directly processed by computers without any need to extract the features first. So far so good, but the main problem raises when it comes to extracting knowledge from such metadata since it requires the highest level of comprehension ability - it requires machines to understand something that was meant only for humans.

To round up the content metadata discussion it can be said, that looking from the personalization perspective very objective fact-based metadata is not as useful as any of the subjective types of information. It is mostly because recommendation system can not extract any meaning directly from such metadata as *Title* or *Author*, unless it uses very static and manual rule-based recommendation which is extremely limited to begin with. If we know such information as the title, the author, the year, etc. we can not say much, if anything, about what the media is like, unless we know the people who created the given media from before and already have formed certain opinion about their work. What content-based recommendation systems can do in this case is to offer "more of the same" by searching for direct overlaps. That leads to very dry and boring recommendations, and due to the lack of novelty, such recommendations are not very useful.

The reality is that people do take such objective metadata (the names of the crew members and music bands) as one of the main criteria to form their expectations about the media. But if we look closer we will see that for people all these names are not just some names in their minds, they are connected with the certain qualities that based on our experience we have come to expect from a certain movie director or a singer. Even with their favorite music bands is

quite often that a listener does not like a few songs, therefore it is not the band's name that makes the listener like certain song or not, it is the inner qualities (inner features) of the song itself. So if we could annotate media not only with some dry objective information, but also with some inner features of the media itself, then we could significantly improve recommendations while satisfying both relativeness and novelty criteria.

It may be worth saying a few sentences again about the novelty aspect. Most people criticize the "more of the same" type of recommendations (usually these are content-based), and for a good reason - it does not introduce us to new things, does not encourage discovering new tastes and broadening our horizons, and in most cases is just too predictable and plain boring. I would argue that people do want "more of the same", it is just that there is a huge difference between "the same" inner features of the media that are meaningful, and "the same" features which do not really mean anything by themselves. Imagine that user's preferences for music are *"emotional songs with complex lyrics with strong focus on sad and mellow moods"*. If we take such preference and return "more of the same" type of recommendations, then we are not bound to any specific author, music period or even genre - we can get all kinds of artists ranging from the *pop* or *jazz* and all the way to *doom metal*. Or if we take this even further and say that the user simply likes the music that is *alive* – then the range and variety of recommendation would cross every single boundary while still satisfying the given preference. In these two examples user would not be stuck in the same genre, the same author or the same keywords, but instead she would stick to the the same underlying meaning. And if we still want more novelty than that, we can always add collaborative-based techniques to the mix, which are relatively easy to implement and are completely domain agnostic. In fact, it is very hard to beat collaborative filtering when it comes down to the novelty of recommendations, but what collaborative filtering lacks is meaningfulness of recommendations and this is precisely where metadata-based, or content-based, filtering techniques have a clear advantage and can contribute.

3.3 Metadata applied

In the overall media personalization picture content metadata serves as a part of the input – the features. The main criteria that metadata has to meet is that it needs to contribute as much as possible to making the input richer so that recommendation algorithms could extract semantic meaning from it. The relationship between metadata and recommendation system is mutual – in order to produce better metadata we need to know how recommendation process works, and in order to optimize the recommendation process we need to know

the metadata and what can be done with it.

Before choosing which metadata specification to apply, first one needs to look into how well it describes the media in question (the media purpose, intended audience, format, information vc. emotions balance, etc.) and secondly one should evaluate how expensive it will be to get such metadata. It is widely considered that the quality metadata is very expensive. In most cases this is true. Basically every time we need someone to sit down and annotate content, the price goes skyrocketing. Therefore people usually try to use all the metadata that is already there before going in and annotating media from scratch. With the professional media there is a lot of editorial and technical metadata that is created in the process of making the media. The most valuable content metadata can sometimes be extracted from other types of metadata [4] although the quality of such content metadata is questionable since there are limitations as to how much content-related information can be inferred.

In the last few years, we have witnessed the appearance and rapid growth of numerous web 2.0 applications on the Internet where people are given a chance to express and share their opinions about all kind of media. This is also metadata. In fact, this is extremely valuable metadata since it gives insights into what common users perceive as being important. And on top of that it is absolutely free. The only drawback is that it is very chaotic in its raw form and usually can not be used directly but needs to be processed before we can put it to use. What it lacks in its structure and the lack of expertise, it makes up for it by its vast size and availability. Therefore every social network is becoming a valuable resource for the metadata that instead of aiming to replace the professional annotations adds significant amount of value.

Since most of the metadata is primarily targeted towards the machines to help them to understand media and to enable automatic media processing, it is absolutely crucial to ensure the use of common language for adding metadata, or to put it another way, to markup content with extra information. For this purpose we have the XML (eXtensible Markup Language) family of languages that is both very wide spread and flexible. XML is a W3C specification since February 10th 1998, but its history predates even the Internet itself. The roots of XML can be traced back to another markup language SGML (Standardized Generic Markup Language) developed in 1980, which in turn builds on GML (Generic Markup Language) from 1969. Today there are over a hundred different

[4] One of the case where this is a widely used practice is the area of TV and Radio broadcast. For example , lots of BBC content metadata come from their content management system, where they have lots of information stating the departments responsible for making certain programs, their purpose, intended audience, etc. Even though it is limited how much we can extract descriptive content metadata from such content management system, but it is there already, and is definitely better than nothing.

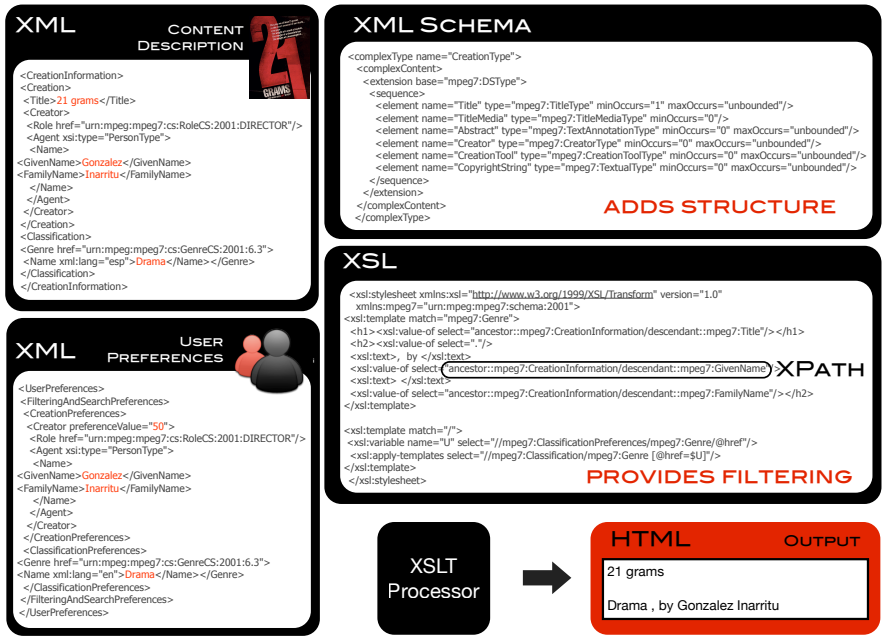


Figure 3.1: An example of MPEG-7 content description containing one movie and user preferences, both in XML. XML Schema controls the structure of content description file, while XSL sets the rules for transformation using XPath expressions and eventually the output is produced in a chosen format (in this case it is HTML)

markup languages based on XML. Here is a list of the members of XML family that are used in media personalization systems.

- XML – the universal format for structured documents and data on the Web, W3C Recommendation since February 1998
- XML Schema – express shared vocabularies for defining the semantics of XML documents, W3C Recommendation since May 2001
- XSL – a language for transforming XML documents, W3C Recommendation since November 1999
- XPath – a language for addressing parts of an XML document, used by XSLT, W3C Recommendation since November 1999

Latter in the thesis three different cases are presented: TV Genres, TV Synopsis and music lyrics. XML is the serialization format in all of them. XML Schema does not appear as such in this thesis but it is an essential part of the whole process working in the background. On top of that many specifications are published in the form of XML Schema (for example MPEG-7 and TV-Anytime) since the schema outlines what are the rules that metadata has to follow. When it comes down to processing the XML files, there is a special functional programming language called XSLT designed just for that purpose, witch uses XPath as a syntax to navigate through the XML document. In this thesis the newest version of both XSLT and XPath are used (XSLT 2.0 and XPath 2.0). XSLT stylesheets were used in this thesis to extract and process TV-Anytime Genre metadata, the process of which is very much alike the example presented in a figure 3.1.

3.4 Dublin Core and MPEG-7

When trying to give state of the art in audio and video metadata standards, it would be useful to start with the two very well known standards that will help to illustrate the scene of media metadata and set the stage for other, more specialized, specifications. These two are the Dublin Core and MPEG-7.

One of the earliest metadata standards (not only in video but also for any other types of resources) is known as the Dublin Core. It was first proposed in the workshop sponsored by OCLC and the National Center for Supercomputing Applications (NCSA) in 1995 [NISO, 2004]. The name "Dublin" refers to the fact that the workshop was held in Dublin, Ohio, and "Core" reflects that this standard proposes only a core set of elements and is expandable. The initial goal was to create a standard for the authors to annotate their own work on the Internet.

Dublin Core standard has two levels: Simple and Qualified [Hillmann, 2005]. Simple Dublin Core Metadata Element Set (DCMES) has 15 metadata elements: *Title*, *Creator*, *Subject*, *Description*, *Publisher*, *Contributor*, *Date*, *Type*, *Format*, *Identifier*, *Source*, *Language*, *Relation*, *Coverage* and *Rights*. The elements can be repeated many times and can appear in any order. Qualified Dublin Core added three additional elements (Audience, Provenance and RightsHolder), as well as a group of element refinements (also called qualifiers) [5].

[5]For example, element *Type* was given a 2nd level containing 12 sub-elements: *Type Collection*, *Dataset*, *Event*, *Image*, *InteractiveResource*, *MovingImage*, *PhysicalObject*, *Service*, *Software*, *Sound*, *StillImage* and *Text*.

Dublin Core is a very high level standard with very flat structure and is extremely flexible. This explains why originally developed with the aim to describe document-like objects (because text resources are fairly well understood), it can be applied to other resources as well. From media personalization perspective the most interesting metadata element is Description. According to the official definition "Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource" [DCMI, 2008].

The real value of Dublin Core comes out of its simplicity and flexibility. Dublin Core is also fully compatible with RDF (Resource Definition Framework)[WC3, 2004] witch is one of the main building blocks of emerging Semantic Web, thus ensuring that Dublin Core has good chances to remain present in the media annotation world.

When people talk about media annotation it is rare that they do not mention one of the biggest standards for media descriptions MPEG-7. Even though in this thesis MPEG-7 is not used directly but it is worth mentioning for several reasons. First of all, it is an example of a huge an complex media description standard also showing that all media problem can not be solved by such an approach. Secondly, parts of MPEG-7 are used in the much lighter TV-Anytime specification which is widely used for TV content descriptions and is used in chapter 5 and 6.

MPEG-7, formally named "Multimedia Content Description Interface" [ISO/IEC, 2002, Martinez et al., 2002], is a standard for describing the multimedia content data that supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer code. MPEG-7 was developed by the International Standardization Organization (ISO), the one which created other standards of the MPEG family (MPEG-1, MPEG-2 and MPEG-4) dealing with the actual coding.

MPEG-7 is a huge and complex standard for describing multimedia. The specification for multimedia description schemes alone is 744 pages. It gives tools to describe multimedia on a number of different levels ranging from low level features (color, texture, sound, shape, motion, spatio-temporal locators, etc) to some higher level semantic descriptions. As it is the case with the most of the released standards that concentrate on being very deep, MPEG-7 offers much more than it is being used by real world applications. An example of content and user preferences annotated with MPEG-7 is given in figure 3.1.

While MPEG-7 looks like it is aiming to solve all the worlds problems and has a perfect vision of how things should work, the reality is a bit different. Today the most successful media description standards are much more light-weighted

compared to MPEG-7 and may not have its depth, but they compensate that by being easy to use and flexible [Tzouvaras et al., 2007, Hendler, 2007].

3.5 Audio metadata

The main scope of this thesis is to contribute to the areas of personalization of complex content types - audio and video. Therefore metadata specifications for images, events, text documents and other types of media will not be discussed leaving more space for audio and video metadata.

After the MP3 music coding standard was created, one of the main problems turned out to be the storing of the information about the file. At the start of their existence the MP3 specification did not have any methods of doing this, thus in 1996 Eric Kemp came up with an idea to add a small piece of information to the audio file containing precisely such information about the item. The standard was called ID3v1 which did not take long to become the de-facto standard for storing metadata in MP3 files. Original ID3v1 tag was of the size of 128 bytes starting with the string "TAG" . The tag had allocated 30 bytes for the Title, Artist, Album, and a Comment, 4 bytes for the year information, and one byte for identification of the genre using a predefined list of 80 different music genres (the list was latter extended by Winamp to contain 148 genres). In 1997 Michael Mutschler made one modification to the existing ID3v1, he trimmed the Comment field by 2 bytes and used this space to store track number information. Such ID3 tags were called ID3v1.1. It shows that back then people did not consider information like Comments to be of much importance, and as we will see latter, such information can be extremely valuable once properly used. A year latter, in 1998, ID3v2 was created. And even though it is still called ID3, it does not share much with the version 1 and 1.1. ID3v2 allows tags to be of variable size usually occurring at the start of the audio file. Interesting point is that ID3v2 allows much up to 256 MB of space for tags in one music file. Tags themselves in ID3v2 are stored in frames, current version allowing 84 frames. The biggest critique for this standard is that not all metadata is extracted from the frames using the same algorithms, that leads to needing dozens of sub-parsers to ensure that we get all the information out (for instance iTunes seems to have problems with the lyrics metadata). Today the specification is being further developed with the most popular version being 2.3 (even though this is not the latest version available).

Even though ID3 is the most dominant music description standard in the market today, there are several others competing for attention. One of them is called OGG Vorbis metadata. It calls its fields "comments", and supports metadata

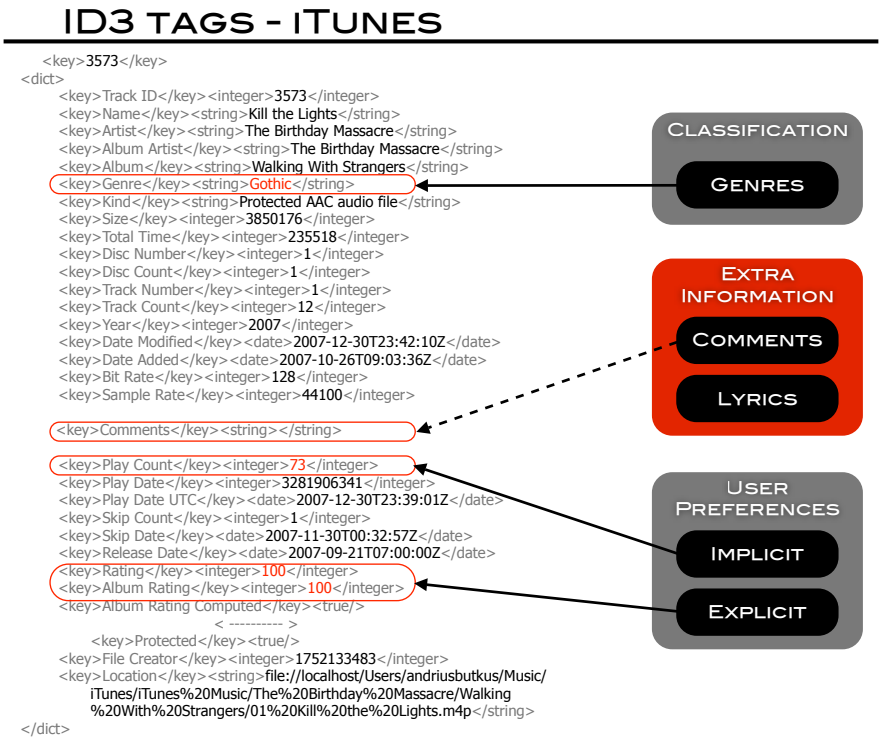


Figure 3.2: Metadata of a song “Kill The Lights” by “The Birthday Massacre” annotated using ID3 Tags

tags similar to those implemented in the ID3 standard for MP3. The other one is called APE and can also be embedded into MP3.

What ID3, OGG Vorbis and APE specifications have in common is that they are based on *key-value* pairs (for example: key ⇒ *Title*, value ⇒ *Kill The Lights*) and do have very flat structure. An example of an audio file described using ID3 Tags is given in a figure 3.2.

ID3 music metadata usually contains such trivial information as the the title of the song <Name>, the artist <Artist>, title of the album <Album>, year of release <Year>, etc. This information is necessary for identification of the song and for classification according to such creation-related parameters. The most common tag that actually tries to describe how the song sounds is the <Genre>. But in ID3v2 user can freely modify its value therefore <Genre> field serves more like a single <Keyword> rather than a true genre which usually comes from

a controlled set (by the way ID3 songs usually do not come with any keywords as such). For the personalization much more valuable are a short field left for the textual description of the song - <Comments>, although it is extremely rare that this information is present. This is because so far the only way to add meaningful comments is to do it manually and is terribly inefficient. An ideal case would be for <Comments> to contain a short review of the song (or even album). Such reviews much like lots of other information can be obtained on the Internet for free. It was previously mentioned in this chapter that lyrics are very special and valuable type of metadata, ID3 gives a possibility to manually add lyrics, but in the vast majority of cases this information is simply not there.

Even though ID3 does not have a special user side since it concentrates only on annotating content, it has several fields that are used to infer user's opinion about a song of an album either explicitly or implicitly. <Rating> and <Album Rating> allows the user to rate items she likes, whereas <Play Count> shows how many times the track was played which gives an approximation of how much the user likes the song.

So far this is all the metadata that is usually available for the client side music player like iTunes or Winamp. I must add that none of the current players offer any intelligent filtering or recommendation solutions. Since players like iTunes usually work offline, they can only offer content-based filtering, since there are no other users to provide information needed to build ratings matrix. And since content-based filtering relies on the informative metadata iTunes only goes as far as building "smart playlist", which are in fact anything but smart, offering nothing more than user defined rules-based filtering on one or more available metadata fields. Those music systems that work online and have a user base therefore can offer collaborative recommendations (Last.fm, Soundpedia, tuneDNA, etc.)

3.6 Video metadata

The necessity to represent video media using metadata has been recognized for quite some time. The need was especially triggered by the appearance of the digital video content on the Internet and on TV [Wactlar and Christel, 2002]. Since user generated content is discussed in the next section, here I would like to focus on the professional video content and metadata standards used to describe it. Most of the professional video content today are still shown on the TV (see chapter 1.1). There are a few specifications created for that particular platform.

One of the earliest video metadata implementations was DVB-SI (Digital Video

Broadcast - Service Information), which is an integral part of the digitalization of television in Europe and other regions of the world. Within the DVB standard family there is a standard specifically for metadata (ETSI EN 300 468) [ETSI, 2003]. The metadata for DVB is also referred to as the Service Information (SI). The standard allows for SI to be inserted right into the broadcast stream and to provide the user's TV equipment (usually coupled with set-top box) with the Electronic Service Guide (ESG) allowing viewers to navigate through the ever growing amount of content.

DVB-SI specification defines a number of tables which enable the DVB service delivery and consumption. There are four PSI (Program Specific Information) tables whose main purpose is to bind all the elements of the transport streams together (specified by MPEG). There are also ten SI tables (specified by DVB) which extend the binding of information allowing the provision of the Electronic Service Guide (ESG).

Within DVB-SI, the Event Information Table (EIT) is especially important as a means of communicating program (event) information. The EIT's can be used to give information such as the program title, start time, duration, a description and parental rating. It is also possible to classify programs using what are known as content descriptors. These are drawn from a two-level DVB-SI genre list.

For DVB compliance it is mandatory to provide information on the currently running program and on the next program. This is the as far as most of broadcasters go. Programs are often not classified using content descriptors and information is only provided on the present and following program/event. This is far from enough to reach any kind of personalization (not to mention the meaningful one). Therefore many broadcasters are either looking for possibilities to extend their DVB-SI with metadata from other standards like MPEG-7 or TV-Anytime (e.g. Danmarks Radio), or they are already using standards like TV-Anytime (e.g. BBC).

Since MPEG-7 was already introduced and since TV-Anytime (TVA) can be viewed as a practical implementation of MPEG-7 lets get straight into the TVA standard. TVA [ETSI, 2004a] is an industry driven specification initially meant for describing and delivering broadcast media to PVR (Personal Video Recorder) devices. With the release of TVA Phase 2 specification [ETSI, 2006] the focus has shifted to enable the delivery and management of all kinds o multimedia content for a variety devices on a number of platforms [Butkus, 2006]. To prove that TVA is not only for traditional broadcast it is worth pointing out that it has been chosen as a standard for DVB-H (Digital Video Broadcast for Handhelds) services [Faria et al., 2006].

Being based on the MPEG-7 standard TVA shares many of MPEG-7 features.

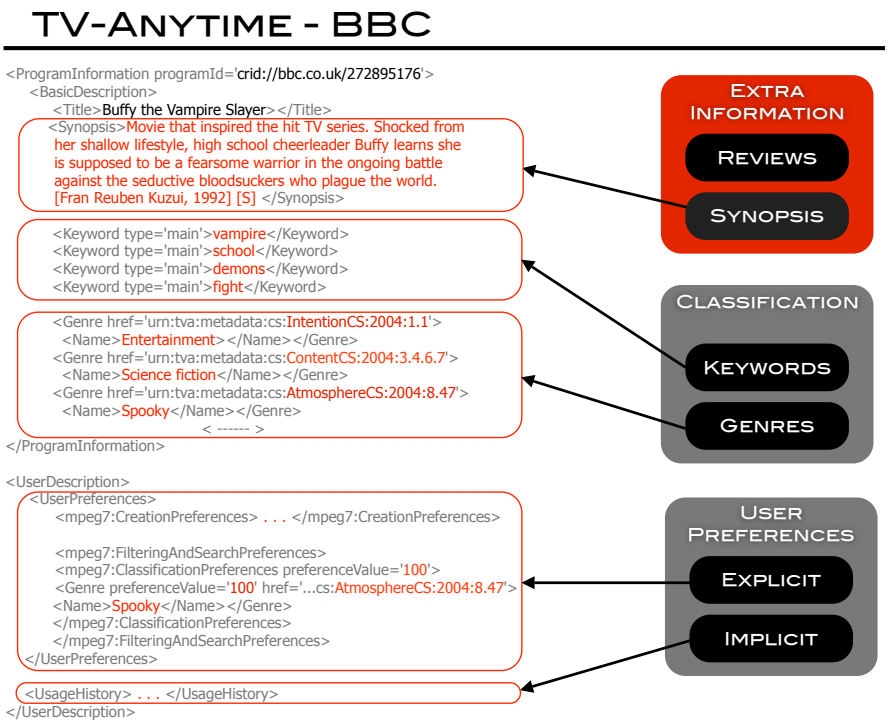


Figure 3.3: An example of a TV program "Buffy the vampire slayer" annotated using TV-Anytime (BBC).

Therefore, on the metadata side, it looks quite similar as the earlier example of MPEG-7 (Figure 3.1). In term of personalization the most important parts of the specification are the ones that describe the content itself and the users. TVA has taken the basic structure from MPEG-7, but added its own Classification Schemes (CS) defining <Genre> information. As for the user part - TVA has borrowed the <User preferences> and <Usage History> blocks directly from MPEG-7. Content related metadata that BBC uses to describe TV programs is shown in the figure 3.3. It is relatively rich description compared with other implementations. In fact, it is almost as rich as it can get without receiving any direct input from the users. Looking at the TVA program description, the user gets the most information from the three metadata elements <Synopsis>, <Keywords> and <Genres>.

<Synopsis> element is the main source of information for human viewers since it is composed using normal speech and usually describes either a plot of the video or the underlying theme (Figure 3.3). TV producers have made synopsis

composing into the form of an art, having teams of people specializing to write interesting and gripping synopsis for this or that channel. The main idea of synopsis is to to serve as an attractor to make users interested into whatever program it describes. Sometimes we may not be able to put a finger on it, but it definitely is able to seduce people into watching the program. We can credit this to our brain which processes synopsis and decides immediately if the show raises some interest for us or not. Machines are not that gifted in this area, therefore it is very hard for a machine to make out a meaning out of a few sentences because it does not help much to simply extract the keywords out of the synopsis directly because what matters is overall meaning and how it makes us feel rather than some stand-alone words. So instead of extracting the words directly from the synopsis, it may produce better results if we use one of the machine learning techniques (for example LSA) to process the text because then we could take words and sentences into context and get closer the overall theme or mood. Such approach is presented in the chapter 6.

<Keyword> metadata is not so special in TVA case and is probably the least descriptive. The only advantage that keywords have against genres in TVA is that there are no restrictions as to which words can be used as keywords whereas genre terms come from a strictly organized taxonomies. <Genre> metadata is the most unique part of the whole TVA content descriptions because we can meet the all the other information in a numerous places elsewhere, while TVA genres are only used in TVA (in their exact form). Following MPEG-7, TVA has proposed to split the genres into multiple categories each reflecting a different aspect of media. In the first phase TVA specification contained 16 different genre taxonomies - Classification Schemes.

For the media personalization the most important schemes are the ones that describe the content inner artistic qualities, for example AtmosphereCS or IntentionCS, whereas some other schemes focus more on the context either of describing either the acquisition or consumption options, for example AudioPurposeCS or PurchaseTypeCS. The most advanced current implementation of TVA is used by BBC, who also made all their metadata publicly available since 2005 under the project *backstage.bbc.co.uk*. BBC chose to use only five of the 16 CS: ContentCS, IntentionCS, FormatCS, AtmosphereCS and IntendedAudienceCS [6]. *Chapter 5* explores the usefulness and potential of using TVA <Genre> for media personalization.

[6] OriginationCS has been found being used in several cases but it is so extremely rare, that it becomes hardly of any use

These are the 16 TVA Classification Schemes from the Phase 1:

- *HowRelatedCS* - Trailer, Alternative, For more information, Recap, etc.
- *TVARoleCS* - Dubber, Puppeteer, Costume designer, etc.
- *RoleCS* - Actor, Reporter, Director, Narrator, etc.
- *IntentionCS* - Entertain, Educate, Inform, etc.
- *FormatCS* - Documentary, Lecture, Show, Representation/Play, etc.
- *ContentCS* - Non-fiction, News, Comedy, Drama, Sports, Basketball, etc.
- *ContentCommercialCS* - Beverages, Leasing, Furniture, etc.
- *OriginationCS* - Live, In studio, On location, Edited, Online, etc.
- *IntendedAudienceCS* - Single, Children 0-3, Professionals, etc.
- *LanguageCS* - English, Russian, Danish, Lithuanian, etc.
- *ContentAlertCS* - Nudity, Offensive language, Violence, etc.
- *MediaTypeCS* - Video, Non linear, Multimedia, Data, etc.
- *AtmosphereCS* - Crazy, Exciting, Sad, Stylish, Heart rending, etc.
- *AudioPurposeCS* - Visually impaired, Directors comments, etc.
- *PurchaseTypeCS* - Forever, For Period, Play Counts.
- *UnitTypeCS* - Day, Month, Year, etc.

Since TVA specification is serialized using XML it is easily extendable. Therefore most of the research in this area has been connected with proposing extensions to TVA. TV personalization problem from recommendation system perspective, was analyzed in [Sullivan et al., 2004] suggesting to use partial overlaps between items to mine more similarity knowledge that can enhance collaborative recommendations. In the iFanzy recommender system the proposed TVA features are implemented to define item similarity based on a set of preferred channels, as well as being used to build collaborative filtering based on usage history to match the user to a stereotype group of other users with the same interests and viewing behavior [Akkermans et al., 2006]. In another content-based approach the TVA metadata attributes have been assembled in a hierarchical user model mirroring a taxonomy of TV program genres reflecting the features of the consumed media [Pogacnik et al., 2005]. Less emphasis seems to have been directed towards how the features may complement each other. Most recent research in

this area (in respect to TVA) has been done by Andrius Butkus and Michael Petersen, where it was suggested to assess the potential for increasing item similarity knowledge by implementing multiple TVA domain specific attributes in parallel and thus extend the semantic dimensionality beyond traditional content genre hierarchies [Butkus and Petersen, 2007]. The potential of using synopsis metadata to extend traditional media classification is discussed in by the same authors in [Petersen and Butkus, 2008a, Petersen and Butkus, 2008b].

User side metadata in TVA is much more advanced compared what is allowed in ID3, since the user preferences mirror the content descriptions [7]. In TVA users are allowed to express their preference (rating) on every single element that is used to describe content. Coupled with <Usage History> information <User Preferences> become a powerful representation tool for users and their needs related to media. One of the main limitations is that usually such user information remains "trapped" in user's set-top box and thus is not be combined with similar data from other users. Another big limitation in TVA approach is that all content metadata comes from the provider (usually the broadcaster (BBC), but is technically open for any 3rd party which, for instance, is capable to provide more elaborate content descriptions) and does not reflect what common users think about the content. This is the main disadvantage of any controlled metadata versus the user generated metadata. BBC and other broadcasters have to spend a lot of time and resources trying to describe the content as best as they can instead of just opening the gates and letting users describe content themselves.

3.7 Folksonomies

As important as it is, the professional video and audio content are not the only media available for the users. Since the emergence of cheap devices capable to record video, that nowadays are also a default feature of any new mobile phone or personal computer, the Internet has become flooded with tons of user generated video content. In the portals like "YouTube" or "Broadcaster" we can find anything from videos of the family picnic to illegally ripped TV shows put online. User generated audio content, while not as popular as the video, is mostly represented with a variety of audio podcasts and amateur made songs. One could argue that the fact that there is more video than audio is partly because video

[7] The fact that the user side metadata mirrors the content side media descriptions means that the complexity of the user description depend on how complex the content side is. The more we know about the content – the more we can tell about the user, since user can express hers opinion about certain content items, but it is the descriptions of the items themselves that will enable to interpret such user action at a much higher level

cameras have become a commodity, and since video usually includes sound (that also applies to music) when people have to choose whether to go for an *audio only* or for *video plus audio*, they go for the latter one. The main point is that if we have user generated content then it does not depend much (if at all) what type of content it is, since in such user generated media the preferred annotation method is *tags* (user generated keywords) forming the basis for folksonomies. In this thesis the folksonomy of the Last.fm music social network is used (see *Chapter 6*) as a source of metadata that can be used to extract similarities between media items thus improving personalization. The goal of this section is to introduce the main idea of folksonomies.

The way folksonomies work is by enabling every user to contribute to the process of describing content. And since the goal is to keep it simple and accessible for everybody (not just the experts) the metadata is basically the collection of keywords. Such keywords have no restrictions in their format or the total number of keywords used. This is as flexible and simple as it can be - the user does not need to worry about whether or not the keywords are compatible, allowed or well formed. Every keyword is a good one as long as it makes sense to the individual user. If we gather all the single user tags into one pile, we get a very personal taxonomy of keywords. It is sometimes referred to as a "personomy". If we do the same thing with all the users in the system, then we get a taxonomy of keywords called "folksonomy". The word "folksonomy", blending the words "taxonomy" and "folk", stands for conceptual structures created by the users themselves. Folksonomies are thus a bottom-up approach whereas standards like MPEG-7 or TVA are top-down approaches. Folksonomies are viewed as a complement to more formalized Semantic Web technologies, as they rely on more latent and emergent semantics [Staab et al., 2002]. One of the key differences of folksonomies compared with formal metadata approaches is that all the intelligence is inside the system hidden from the users, therefore it does not require any professional knowledge to be able to annotate media [Hotho et al., 2006]. One could classify all media folksonomies into two groups: the one where users share their own amateur content (YouTube) and the ones where they only share their opinions about professional content created by professional users (music bands, movie studios, etc) (Last.fm). Social networks that rely mostly on user generated media usually have much greater variety of content compared to the ones where user takes place more as a critic and not as the creator. But on the other hand, the folksonomies based on user sharing not the content itself but their opinions about it have potentially better correlation between the users since the average user base for every individual media item is generally larger.

Since there is no artificial structure that users who tag the content must follow, what we get as a result is a bag of keywords. As different as people are in terms of their taste we are very much alike when it comes down to how we think and perceive things on a very basic level. We share a great deal of common

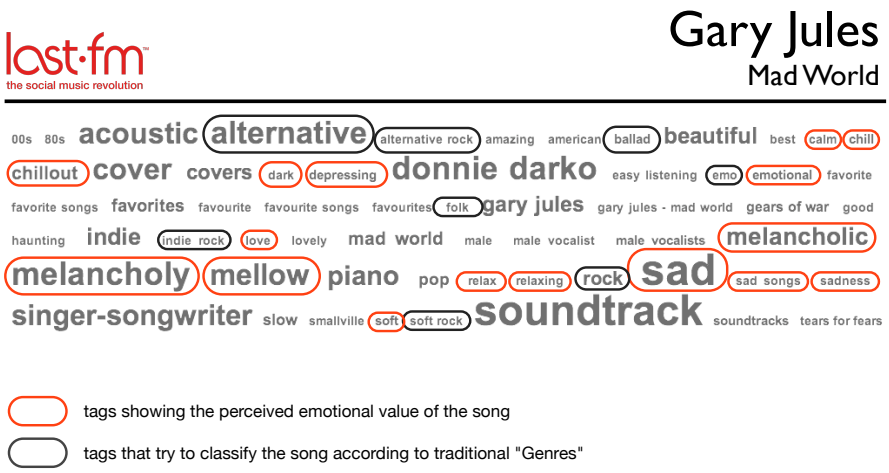


Figure 3.4: The tag cloud for the song "Mad World" by Gary Jules (source: Last.fm)

knowledge and common experience, on top of that there are certain cognitive laws that apply to all of us whether we are aware of them or not. All this leads to the inevitability that there will be some overlap between the tags that different people assign to the same piece of media. Repeating tags are not a source of noise since the goal is not to capture the greatest possible number unique tags for a certain piece of media, but to capture the tags that express the way the the media is perceived. Therefore if a certain tag appears in a media description many times, it shows that people agree on this particular quality of a song or a movie. This is very valuable information since it helps to form a ground truth for further recommendations [8]. When it comes down to visually represent tags in a folksonomy the most popular way it the "tagcloud" (Figure 3.4). From the first look, the the tag cloud looks like a simple collection of tags, but on top of that it also shows the relative frequency of each tag by variation of different font sizes (the more frequent the tags appears to be – the bigger the font size) [Sinclair and Cardew-Hall, 2008].

The idea to use tags for content annotation is not particularly new or revolutionary: keyword-based retrieval has been around for a while in the area of information retrieval. From the processing (recommendation systems) perspective such tags can be viewed in two different ways. One way is to claim that tags, in contrast to the formal semantics provided by the Semantic Web standards, have no semantic relations, and are just flat collections of keywords

[8] more elaborate discussion on the ground truth and how people tend to agree on certain features is presented in the *Chapter 4*

[Tzouvaras et al., 2007]. Another way would be to recognize that there are millions of users performing the tagging and when we have a scale of such size then the "size matters" effect kicks in and tags start to form patterns enabling us to infer a lot of structure information and get closer to predicting the inner meaning of the media that is represented by such tag cloud.

If we take a tag cloud of a song as an example (Figure 3.4) we might initially see it as a random collection of words. But if we look closer we see that there are certain groups of different terms aiming to express different things. For the accuracy needed in this thesis it is enough to identify two groups of terms in a folksonomy. First group of terms is very much related to the classification of music and looks very similar to what people would normally call a *genre*. This relates to a style of music without any explicit comments about how it actually sounds. Just because two songs belong to a *Alternative Rock* genre, it does not mean that they will be perceived as being similar. On top of that, as it was discussed previously, the ground truth for genres is very vague since everybody tends to have their own understanding of what genres (for example *Rock*, *Indie Rock* or *Alternative Rock*) mean for them. In the figure 3.4 the tags related to the traditional genres are circled in black, while the tags related to the emotional context of the song are highlighted in red.

Another big group of tags are related to the emotional context of the song, rather than classification. Tags like *Sad*, *Dark* or *Mellow* show the emotional reaction that the song raises in people. As personal as we may think such emotional responses are, sampled over a large user base they can reveal the general emotional context of the song. As can be seen in the figure, the most common tags for the song by Gary Jules *"Mad World"* are the *Soundtrack* and *Sad*. Even though some people consider *Soundtrack* to be a genre (and in principal it can be called a genre since the very definition of a genre is quite vague and not clear) but at least we can agree that such genre does not give us any information about the song other than the fact that it has been used in a movie. *Sad* on the other hand is very informative tag and gives us a good idea of what we can expect from the song. After those two, the other frequent tags are: *Alternative*, *Melancholy*, *Mellow* and *Donnie Darko*. *Donnie Darko* tag shows the name of the movie where this song is used as a soundtrack (for the people who have not seen the movie, it bears no information whatsoever), *Alternative* is the genre related tag that can refer to many things, and while being more descriptive than *Soundtrack* it still is relatively vague. Two emotional tags that come out as a frequent ones - *Melancholy*, *Mellow* - attempt to describe the emotional context of the song and coupled with the tag *Sad* give clearer indication of what this song sounds like compared to the information we get from tags *Alternative* and *Soundtrack*. In the case of this particular song we can see that emotional tags outnumber the genre-related tags. This is not always the case, but on average we would have approximately even amounts of

emotional and genre tags.

The rest of the tags found in a folksonomy any neither genre nor emotional. These are usually words attempting to classify content to a category that usually does not cary much information about the song itself, but tells something about the authors (gary jules), locations (american), time period(00s, 80s), etc.

Looking from the economical perspective, folksonomies are a source of cheap metadata [Basu et al., 1998]. In fact "cheap" in most cases even becomes "free" metadata since all the costs associated with the annotation falls on the users of the system. From the user point of view such costs are basically time spend tagging media, but if the users feel that they are getting some kind of value back and that value is greater than the cost (time spent) then it is a win-win situation where both users and the service providing entity are satisfied. One could discuss a lot what kind of value the users are receiving. But to put it in a short way, most value comes from two areas: social relations and discovering new products. As the name suggests, the social networks build on one of the basic human needs to be socially active. While the Internet has eliminated the necessity to be in the same geographical area, the social network portals are the glue that help to connect the like-minded people all over the world. Another benefit for the users is that social networks enable not only to share their media taste with other people and to meet them, but it also helps users to find new content, the one that they otherwise would not be aware of. Every social network (both in the real and online worlds) acts as a recommendation system. If nothing extra is done to endorse more complex recommendations, then the user is still exposed to one-on-one type of recommendations from the other users. But in the case of media, most social networks try to combine recommendations and opinions of all the users - collaborative filtering, and in some cases such recommendations are also enhanced with the content-based component depending on how much content metadata is available and if any advanced filtering techniques are used.

3.8 Conclusions

The first three chapters of the thesis have presented the state of the art of media personalization looking from three different perspectives - economical, recommender systems and media metadata. This is where we are at the moment. Memory-based collaborative filtering is the most popular approach and mostly relies on implicit preferences derived from the usage history. More advanced cases also allow users to express their explicit feedback by rating media items. If this is all that is present, then the performance of such system depend on the efficiency of the collaborative filtering algorithms which might have reached

the upper limits of what can be done with very limited user data that they are presented with. Biggest challenge still left to be solved for the collaborative filtering is the scalability which is the core issue for every practical implementation. Scalability aside, there is nothing much that can be done to significantly improve the quality of the recommendations that are based purely on collaborative filtering when none of the content features are known. But the truth is that at least some of media features are always known, even if it is as limited as the title or the artist information (not to mention editorial and technical metadata). Utilizing such information is a job of the content based filtering. Therefore most of the media recommendation system implementations today have both the usage history data and content metadata, and they use hybrid recommendation algorithms where both collaborative and content-based filtering are in place. Content-based filtering quality depends solely on how much knowledge we can extract from the metadata that we have.

One way to approach this problem is to create very advanced and precise metadata descriptions so that knowledge extraction would be easy. MPEG-7 is a good example of such approach but has proven to be very expensive and way too complex for a common user. Even its quality can be questioned since it comes from a group of experts and quite often it does not reflect how the users feel about certain piece of media.

Another way would be to allow users to annotate media themselves thus eliminating the price factor of metadata. The most popular way for the users to describe media is through tagging it with keywords. While it does not seem like there is much quality there (these are just amateur users), it only begins to make sense when all the user annotations are pulled together forming folksonomies and representing the way media is generally perceived by the users. The main challenge is this case is how to utilize such unstructured user generated metadata to infer knowledge about media items, and through them about the users themselves. To do that it is necessary to understand what are the main principles that work in the background when people categorize or classify media, or any other items in general.

The next chapter presents the cognitive theories and methods that govern our perception of similarity and categories. It also introduces several methods that are used to explain and simulate such human behavior on the machine side.

Chapter 4

Categorization based on Cognitive Principles

Before starting with this chapter lets step back for a while and review the main research question and what it means after the first three chapters. The main overall goal of the thesis is to make recommendations (and thus personalization) more meaningful for the user. At this stage of the thesis we already know that the most popular recommendation systems employ hybrid approaches for recommendations utilizing both collaborative and content-based filtering, where the content-based filtering has not yet been using its full potential. We also know that the quality of this component depends on the quality of the metadata, which refers to how much knowledge we can extract from such metadata. What we want to be able to extract from the metadata is the information that tells what the media item is all about, what is it like and where does it fit.

Since editorial and technical metadata are both very objective and straightforward, the main focus is on the content metadata. To be even more precise, the main focus is on such content metadata that attempts to describe what the media is like, rather that simply stating all the creation related details. As we saw in chapter 3, such "subjective" metadata is used in nearly every metadata standard in the form of keywords, classification schemes or free text descriptions. It is also very popular in the user generated folksonomies forming roughly one third of the tags. The main problem so far is how to use all this metadata to

improve our understanding about media and thus improve recommendations. In order to that it is important to know where all those keywords, genres and classification terms are coming from.

This chapter explores the way people categorize things and explains how all the "subjective" content metadata is formed in our minds. Since the personalization process needs to be automated it is very crucial not only to understand how our brain categorizes things, but also how we can simulate this process on the machine side. This chapter addresses this need and presents several popular machine learning techniques that can be used to simulate categorization of media and are later used in the *Chapter 6*.

4.1 Categorization

In order to cope with the constant flow of information, humans much like any other organisms must group their experiences into meaningful categories. While numerous experiments prove that the categorization of stimulus is essential for cognition, scientists still debate on how exactly our brain performs categorization. This question has been tackled by philosophers, psychologists, mathematicians, linguists and computer scientists among many other fields. One of the most common definition of a category is the following (taken from the Oxford Dictionary of English):

> *Category is a group of objects having particular shared characteristics.*

The most important part of the definition is stating that objects must have "particular shared characteristics". Therefore the main challenge with categorization is determining which characteristics are shared among objects and how important are they. Which leads again to one of the key terms in this thesis - *similarity knowledge*. It is impossible to talk about categorization of something without discussing similarity. In the cognitive sciences, similarity plays an essential role in how humans acquire and categorize information [Spiteri, 2007]. Or as Tversky puts it – similarity is "an organizing principle by which individuals classify objects, form concepts, and make generalizations ... it is employed to explain errors in memory and pattern recognition, and it is central to the analysis of connotative meaning" [Tversky, 1977]. Therefore the question what makes two things appear in the same category, starts with first determining how similar those two things are.

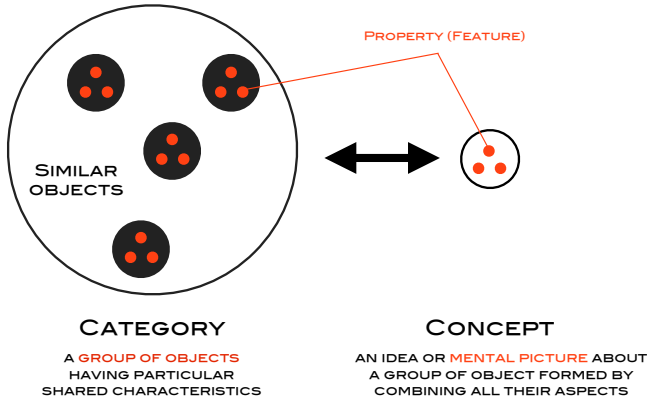


Figure 4.1: Relationship between a category and its mental representation - a concept, all glued by similarity.

Another important term that is used often in cognitive science is a *concept*. It can be understood as a mental representation of a category in our brain that serves as one of the main building blocks of our knowledge. Popular definition of a concept, according to the Oxford Dictionary of English, is the following:

> *Concept is an idea or mental picture about a group of objects formed by combining all their aspects.*

There is a close mutual relationship between categories and concepts. Categorization involves characterizing something by means of concepts so, for example, "my concept of dog allows me to pick out a category of entities that I would call dogs" [Prinz, 2002]. Both categories and concepts are formed based on the similarity knowledge about a group of objects. "Conceptual coherence has been associated traditionally with the notion of similarity, that is, objects, events, or entities form a concept because they are similar to one another" [Spiteri, 2007]. In its most simple form similarity means sharing features 4.1.

How is all this related to media personalization? The idea of personalization is all about automatically finding media that is similar to what a user likes. User preferences in many cases can be inferred from the behavior exhibited in the past - usage history (see chapter 2.1). Therefore it is crucial for the personalization system to able to distinguish similarities between content items and be able to group them into the meaningful categories. The notion of similarity forms the basis for most of the approaches used in the Library and Information Science (LIS), which is one of the fields that modern media personalization builds on:

"classification is, in its simplest statement, the putting together of like things, or more fully described, it is the arranging of things according to likeness and unlikeness" [1] [Richardson, 1964].

Such fundamental questions like *What is similarity?* or *How do we group things together?* has interested humans since the beginning of time. One of the first theories dealing with categories, concepts and similarity comes from the ancient Greece philosophers, particularly from the works of Plato and Aristotle. This is known as the classical model of categorization. Classical model is based on the necessary and sufficient conditions. It means that every category has a number of very clear requirements that the object needs to satisfy in order to be a part of the category. Those requirements are expressed through the various features of the objects. To put it in other words – "categorization is a process of checking to see if the features that are part of a concept are satisfied by the item being categorized" [Spiteri, 2007].

This model first defines very clear boundaries for a category (by formulating the necessary and sufficient conditions) and then simply puts objects either inside the category or outside. This leads to all the members of the same category being equally similar to each other because they possess the same properties. That means that similarity is symmetrical, since what is true for one entity in the category is true also for another [Laurence and Margolis, 1999]. This shows that all the features in this model are considered to be either present or absent - they are binary.

From the media personalization perspective classical categorization model can be seen as filtering based on a very objective metadata following clearly defined rules. For example, imagine that my preferences for movies only state that I like an actress Naomi Watts, and that I want to retrieve only the movies where she appears. Then this condition clearly divides all the movies into two categories - the ones with her, and the ones without. Such personalization is neither challenging nor interesting and most of all it is definitely not novel. On top of that, as it was discussed in chapter 3, the fact that the same actress appears in ten different movies does not say that those ten movies are similar by any other criteria other than then all having the same actress. Such categorization (like a category "All movies with Naomi Watts") is very artificial and does not reflect what people mean when they say that certain movies are similar.

If we want to have more meaningful and personal categories we have to use more complex features that are not so straightforward and obvious. Imagine categories like *Comedy* or *Heavy Metal*. How would we define them using clas-

[1] Classification and categorization are similar terms and even though they do have subtle differences in the scope of this thesis they mean essentially the same thing - grouping things into classes according to their shared characteristics.

sical categorization model? Interesting thing about categories and concepts is that "knowing a concept" is not the same as "knowing a definition of a concept". One might say that it is perfectly enough to know a concept and we do not really need to define it. In this case the process of cognition would seem to have a middle step which could be called *And then something magical happens.* Besides the basic scientific curiosity (what actually happens?) there is an extra motivation for us to figure our the definitions, because we want to automate this process of categorization. And we do need to have some sort of definition if we want to delegate the categorization task to the machines. A good illustration of a concept that we all know but can not easily define is a Wittgenstein's *game* [Wittgenstein, 1953]. It turns out that a *game* is not the only category that is hard to define - it is the case with nearly all of them. For example, everybody knows what a comedy is once they see it, but to define a comedy is quite another challenge and is much more complex as it may appear. Most of the concepts have somewhat fuzzy boundaries and do not follow the classical categorization model. The more scientists started to think about it the more apparent it became that the classical categorization paradigm has severe limitations and the whole idea of concepts and categorization needs to be rethought.

4.2 Prototypes and exemplars

Aristotelian view on concepts and categories has been very influential and held its ground for surprisingly long time - all the way up until the middle of the last century [2]. But as a result of growing evidence and dissatisfaction concerning classical model, new ideas started to appear. One of the most influential of them was the *family resemblance* first proposed by an important twentieth-century philosopher Ludwig Wittgenstein in the Philosophical Investigations [Wittgenstein, 1953]. The central idea was that a category of objects does not have to be defined by one essential common feature, but instead can have numerous similarities which may overlap but are not necessarily shared by all object in the category. In contrast to the classical model, the family resemblance approach does not have any necessary or sufficient conditions. Therefore instead of being either inside or outside of the boundaries, it is a matter of degree of how much something belongs to a category or not. This is called a *graded membership.*

Family resemblance gives us a probabilistic estimation of an object belonging to a certain category. It means that if an object X has features a and b then it probably belongs to a category Y. For example if an object in front of us has wings and is able to fly, then it is probably a bird (it may still be a bat, but there is much smaller chance for that). Notice that there are no clear predefined

[2] for more on a classical model see [Smith and Medin, 1981, Labov, 1973]

boundaries for a category, instead we have a whole range of objects with different probabilities of belonging to a given category. This idea can be translated into a psychological theory in a few ways - prototype and exemplar [Dopkins, 1997].

4.2.1 Prototype model

Family resemblance and graded membership ideas have lead to the creation of a new theory to explain concepts and categorization using prototypes. In the 1970s Eleanor Rosch has proposed a model that was a big shift away from the classical categorization model. She called it a prototype model (also known as a Prototype Theory) [Rosch, 1973, Rosch, 1975, Rosch et al., 1976, Rosch, 1978]. As the name suggests, prototype model assumes that the category can be formed not by defining the borderline but by specifying the center of the category - the prototype.

As its central idea the prototype theory assumes that in every given category certain members are more prototypical than others (for example robins are thought to be more prototypical birds than penguins or emus). Therefore prototype model acknowledges the existence of family resemblance and graded membership since all category members are not equal - some are more prototypical than others.

What do prototypes mean in terms of media personalization? According to this model it means that for every category that we can think of to classify media there is a prototype in our mind and that all the members of the category are different in terms of how close they are to the prototype. We can think of a prototypical drama or a comedy, or on the music side people talk about prototypical pop or heavy metal music, and sometimes even the prototypical music from the 80s. We may argue how precise are such prototypes, but people use them nevertheless, despite the fact that prototype itself depends on what family members we have encountered in our life so far, causing different people to have different prototypes.

One of the main challenges is to define a prototype which can mean a number of things depending on the specific case. Usually prototype can be thought of as an object that has the average features of a given category. For example a prototypical family can be imagined as having an average income, being of average age, etc. Sometimes when we can not talk about averages, then prototype can simply be the most typical instance of a given category. For example the prototypical family will have 2 children instead of the 1.74 given by the statistics. And in some cases prototype can be an object with an ideal set of features, something that other category members can look up to. For example a diet coke

prototype may contain 0 calories even if it turns out practically unachievable.

Another very important question for any categorization model is to explain how categories are formed and how the new members are added to a category. Lets take a category *dog* as an example. First a person must encounter several dogs (in real life, in literature, movies, etc.). After that the person creates a mental image containing all the features that this person thinks are general for dogs. As mentioned before, some features will be averages (like height), some will be most typical ones (like the color). All this makes up a prototype, which is closely related to knowing a concept of a dog[3]. Once we have a prototype, then a category is formed around it with new members being added on basis of resemblance to a prototype.

Every single object always belongs to more than one category because we can use different levels of generalization. For example an object may be a german shepard, a dog, a mammal or simply an animal, all at the same time. Why does it make more sense to talk about prototypical dogs and not prototypical animals? This happens because of generalization and it turns out that certain abstraction levels are more meaningful and informative than others. These levels are known as the basic categories and were introduced in the prototype theory as the basic level that has the highest degree of cue validity [Rosch, 1978]. For example if we are presented a picture of a dog and asked "what is this?" we are most likely to say "this is a dog", compared to a more precise definition "this is a German Shepherd" or more general "this is a mammal" , even though all three answers are correct. According to Rosch the basic level category has greater psychological significance. For media personalization basic categories relate to how we name genres and how we tag content in the social networks.

Basic categories can be defined in several ways. First they can be understood as the highest level at which a single mental image can represent the entire category (dog is a basic category whereas animal in this case is superordinate category). They can also be looked at as the highest level at which category members have similarly perceived overall shapes (dog, but not animal; car, but not vehicle). Or they also be defined as the highest level at which a person uses similar motor actions for interacting with category members (separate motor programs for interacting with chair, bed, table, but not for interacting with furniture) [Reisberg, 2001].

There has been a significant amount of research done to gather evidence for the prototype theory. The most important empirical evidence is listed bellow in the form of five different cognitive test that has been carried mostly by Rosch and

[3]Prototype theory states that to "know" a concept may mean to have some mental representation of the concepts protype.

her colleagues:

- *Sentence verification test.* In this one the participants were presented with various statements and they needed to determine whether the statement is true of false as fast as they can. The speed of decision was taken as an approximation of how prototypical a certain member is. For example a statement "a robin is a bird" has caused people to decide much faster compared to a sentence "a penguin is a bird" [Smith et al., 1974].
- *Production test.* The participants were asked to name as many members of the category. Results have showed that the more prototypical members are being named earlier compared to the less prototypical [Mervis et al., 1976].
- *Picture identification test.* The participant were told that they were about to see a picture that may or may not be a dog and they were asked to hit the "yes" or "no" button as soon as they can. Results have shown that pictures of dogs like the German Shepherd are more quickly identified as dogs compared to some less prototypical dogs such as Chihuahua [Rosch et al., 1976].
- *Explicit judgement test.* In this one people were presented with a number of different members of a category and were asked to rate how prototypical the members are to a given category [Rosch, 1975].
- *Induction test.* The people were told some new facts about a member of a category, and were evaluated on their ability to extrapolate the new information to the other members of the same category. Results have shown that people are much more likely to make inferences from the typical member to the whole category, but will not make inference from an atypical member to the category [Rips, 1975].

Mathematically prototype-based categorization can be visualized using Voronoi tessellation [4]. The main property of the Voronoi tessellation is that if the space is based on the Euclidean metric (that means that we can calculate distances by drawing a straight line between two points), then the tessellation always partitions the space into convex regions (see Figure 4.2). A region is considered convex if for every two points in that region we can draw a straight line and then all the points on the line belong to the same region. The problem is that Voronoi tessellation is only applicable to low dimensional spaces where the distances can be expressed in Euclidian metric, therefore it fails to capture the complexity of the multidimensional reality that is usually the case in the real world.

[4] Voronoi tessellations, also called Voronoi diagrams, are named after Russian mathematician Georgy Fedoseevich Voronoi (1908). Informal use of be traced way back all the way to the "father of modern philosophy", Rene Descartes in 1644.

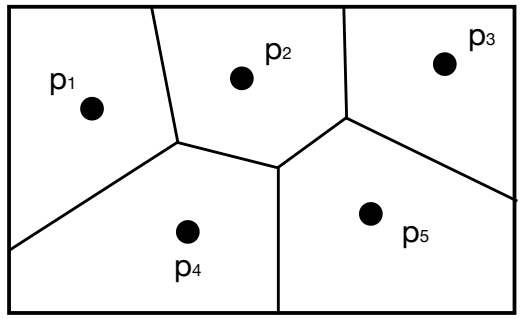


Figure 4.2: An example of Voronoi tessellation of a space based on the position of the prototypes: p.

4.2.2 Exemplar model

Another way to think about typicality effects and family resemblance is through exemplars. The main idea here is that the categorization is done not by comparing subject to one mental prototype of the given category, but comparing it to all the other objects that we have encountered that we know belong to the given category [Nosofsky, 1987, Nosofsky, 1986, Medin and Schaffer, 1978, Brooks, 1978]. This approach overlaps with the prototype theory in a way that here we also compare the new object to something else, the main difference is that the comparison is done between the object and multiple exemplars, rather than the object and one prototype. Exemplars are always real objects whereas a prototype may be an just a collection of prototypical features but may not necessarily represent an existing object. This characteristic allows exemplars to preserve more information about the individual features and their variability compared to a prototype [Reisberg, 2001].

Even though these are two separate models people use both prototypes and exemplars depending on a number of factors. One the important factors deciding which model is applied depends on people's level of expertise in a given field of media. For example, in the beginning users tend to rely mostly of exemplars when drawing their conclusions on whether something belongs to one category or the other. When people get more familiar with the field and gather a number of exemplars, then they are often found to form a prototype of a category based on the exemplars they have encountered. Empirical evidence supporting exemplar model can be found in the works of [Brooks, 1978, Medin and Schaffer, 1978, Rips, 1989] just to name a few examples.

People use exemplars when talking about media as well. For example people

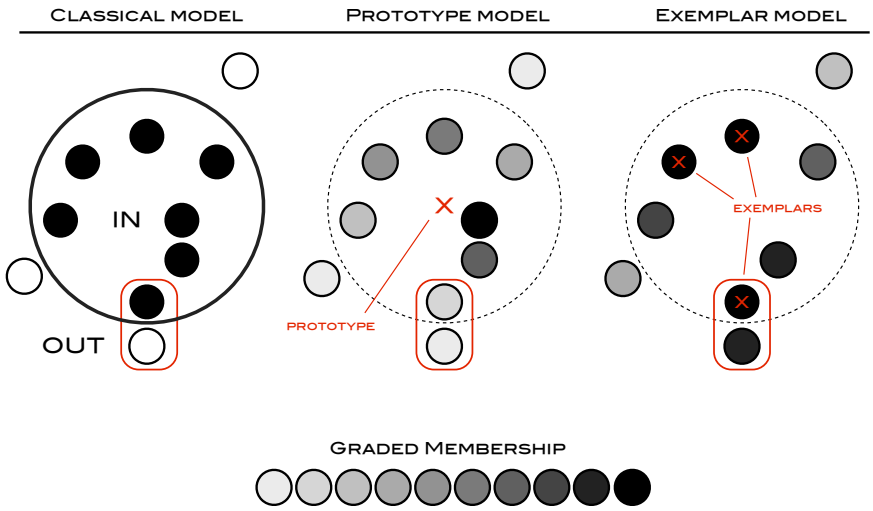


Figure 4.3: Classical, prototype and exemplar models of categorization

tend to describe their taste for music or movies by naming a number of exemplars - music bands, movie titles, etc. This is especially the case with the users who do not know the domain that well yet.

4.2.3 Beyond prototypes and exemplars

So far we have established that typicality effect play a big role in a categorizing process used by both prototype and exemplar models. But despite the amount of evidence for both of the models, there are many cases when we overrule the typicality effect and draw our conclusions on deeper, and thus sometimes not easily observable features. There are a number of cognitive experiments supporting this view. One of the experiments reported people categorizing a whale as a mammal even though it may look very typical to all the other fish [McCloskey and Glucksberg, 1978]. Another experiment used objects that were changed to look very different from their respective prototype or any of the exemplars – imagine a lemon that has been injected with sugar (therefore it is sweet), has been painted with red and white stripes and run over by a car so now it is flat. It does not look anywhere close to a lemon, but people agree that is still a lemon, no matter what [Rips, 1989]. These experiments have suggested that while typicality through prototypes and exemplars alone is good for "quick and dirty" categorization, people use more than that when they categorize objects and build concepts.

Here is a list of what humans normally use for that task [Reisberg, 2001] :

- a prototype for each concept
- a set of rememberable exemplars
- a set of beliefs about concepts
- an understanding how those beliefs fit together

It was stated before that concept is an idea or mental picture about a group of objects formed by combining all their aspects. The most important part of the definition is at the very end of the sentence - "combining all their aspects". This highlights the idea that in order to call something a concept we need to see it from a multiple angles. The next section builds on prototypes and exemplars but approaches categorization from a more complex and multidimensional view where a concept is no longer limited by 2D voronoi tessellation but instead takes shape of a multidimensional space.

4.3 Conceptual spaces

There are a few theories dealing with knowledge representation using a space. The most notable one is called Conceptual Spaces theory and was introduced by Peter Gärdenfors in his book "Conceptual Spaces: The Geometry of Thought" [Gärdenfors, 2000], followed by numerous articles ([Gärdenfors, 2001, Gärdenfors, 2004], etc.). Gärdenfors theory builds on the earlier research on prototypes and exemplars (among other things) and can be seen as a possible implementation of categorization by creating and combining multi-dimensional feature spaces. In this section I would like to present the parts of the Conceptual Spaces theory that deal with concept formation and that can be used to give a cognitive perspective to how we could make sure that we are looking for the similarity in the right place when we are talking about media.

One of the key aspects that Gärdenfors proposed to look at the concepts as the geometrical spaces – "A conceptual space is built upon geometrical structures based on a number of quality dimensions" [Gärdenfors, 2000]. That being said, the some of the immediate questions that come from his statement are *What are the dimensions?*, *How many dimensions do we need?* and *Where do those dimensions come from in the first place?*

There are several new terms introduced here: dimensions, domains, regions, properties and concepts. I will introduce them one by one, going from dimensions all the way to concepts, and then at the end I will present an example that connects everything together.

4.3.1 Dimensions and domains

The dimensions are one of the fundamental building blocks used in the Conceptual Spaces theory, and are understood quite literally. Some of the most obvious examples of dimensions are space (height, width and depth), time, weight, color hue or musical pitch, just to name a few. Every dimension has a topological structure - linear, spacial, binary, tree, etc. (see Figure 4.4). Topology is important when trying to calculate the distance (and thus similarity) between two points in that dimension. Since we are living in a three dimensional world, we are used to perceive the distance between two points as a straight line. But it is not the case in all the topologies, for example in a tree structure, the distance is perceived as the length of the path between two points rather that a straight line. There are also dimensions consisting of binary values where we can be either there or here, but never in between, a good example of this kind of dimension is gender.

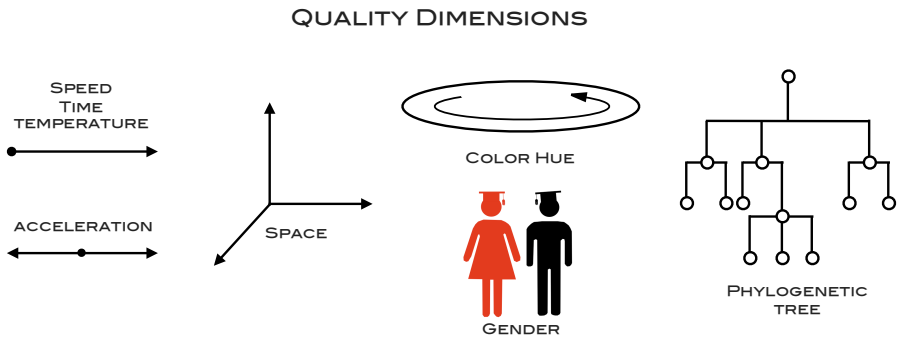


Figure 4.4: A few examples of different quality dimensions having different topological structure.

To answer how many dimensions we need, we need first to look at where they come from. In his definition of a concept Gärdenfors talks not just about any dimensions, but about *quality dimensions*. It is believed that some of the quality dimensions are innate or are developed very early in life. That means that they are in some way hardwired in our brains and mostly are related from the

sensory information (what we can see, hear, taste, etc). But this is only where things start. Dimensions of our world can be learned, and this ability has been confirmed by monitoring children and seeing how they learn to identify new dimensions as they grow [Gärdenfors, 2001]. Supporting this notion, numerous cognitive studies have indicated that there are many of quality dimensions that do not come directly from our sensory information, but are learned through experience and are influenced by many factors, such as social rules or culture [5]. Gärdenfors distinguishes between the psychological and scientific methods to determine the dimensions. He argues that "since the dimensions are cognitive entities their topology should not be determined by scientific theories, but by psychophysical measurements which determine the structure of how our perceptions are represented" [Gärdenfors, 2001]. This is quite the opposite of what we see for example in TV-Anytime or MPEG-7 classification schemes which follow an approach to represent semantics as a taxonomy. To put it in another words it suggests that a choice for dimensions should come from our perceptual system, whether it is based on sensory information or cultural or any other areas, instead of being forced from outside, for example building very detailed yet artificial taxonomies that have no cognitive grounding.

The current knowledge of what are the quality dimensions in many domains is quite limited and therefore this is one of the main challenges for modeling conceptual spaces. There are a few examples where the dimensions are fairly well understood already. Probably the most common example is how we perceive colors. It turns out that our cognitive representation of colors can be described using three dimensions: hue, saturation and brightness. Hue is represented by a traditional color circle and therefore is a circular dimension. Saturation ranges from gray (no color intensity) to greater intensities which of are the most saturated versions of respective colors. As can be seen, saturation is a linear dimension. The third dimension is brightness, which goes from white to black. Combined together the three dimensions form a space, often called a color spindle (see Figure 4.5).

Another example is the way we perceive taste. One of the popular models was proposed by Henning in 1916 (there are more than one model proposed to represent taste). He argued that since the perception of taste come from four different types of receptors (salt, sour, sweet and bitter), the we could represent taste in a four dimensional space forming a tetrahedron shown in figure 4.5.

As can be seen from the two examples above, the dimensions are very fundamental blocks and if we have no idea about them then we can not talk about building concepts. Therefore it is very important to know or at least to assume

[5] For example the dimension of time is perceived differently in different cultures. In some cultures time is circular, in others linear, etc

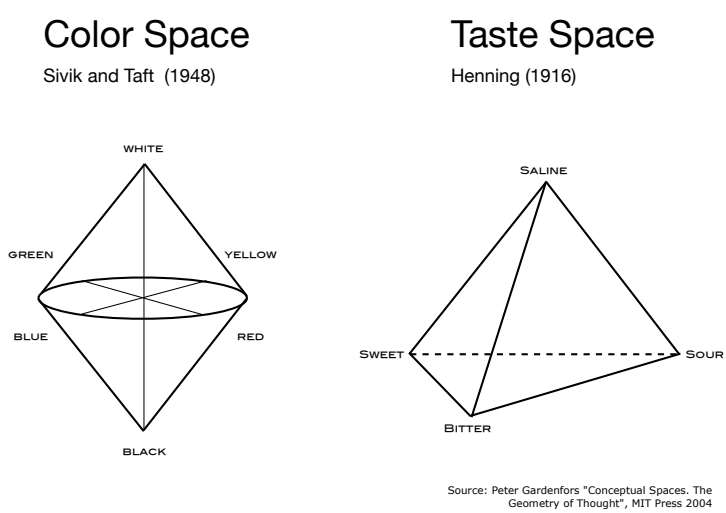


Figure 4.5: Visual representations showing how humans perceive color and taste.

certain dimensions, but maybe even more important is not a dimension alone but how it interacts with other dimensions. This interaction among dimensions can be expressed as the dimensions being integral or separable. Gärdenfors here builds on earlier research in this area [Maddox, 1992] and calls two dimensions being integral if we can not assign an object a value in one dimension without automatically assigning a value in another dimension. For example, if we take the color spindle, we can not have hue without saturation or brightness. Therefore we can say that three color dimensions are integral and they form a color domain. On the other hand there are plenty of cases where an object has dimensions that are not connected and thus are called separable. For example, we can specify the length of an object without specifying its saturation.

If we specify that certain dimensions are integral then we can talk about dimensions forming the domains. Therefore in the previous examples we actually have a color domain, a taste domain (Figure 4.5) because they are build from integral dimensions and form a uniform space.

Once we have the dimensions then in many cases when we can also talk about the distances between points. Once the coordinates of two points are known for all dimensions, then it is possible to calculate the distance between those points using the metric that fits the space. If we reduce the space into a two-dimensional space the we can use Euclidean or Minkowsky metrics. If the dimensions turn out to be integral then the distances in the domain can be calculated using Euclidean metric, whereas if the dimensions are separable then we can

use Minkowsky city block metric. If we stick to the multidimensional space, then we can express an object as a vector in a multidimensional space, which allows us to calculate cosine between two vectors as a measure of similarity (see Chapter 2). The main reason for knowing the distance is because it serves as an approximation of similarity, and similarity is the key aspect of the learning and categorization processes.

In Conceptual Spaces theory the dimensions are used to "assign properties to objects and to specify the relations among them" [Gärdenfors, 2000]. And therefore the conceptual space itself can be defined as a set of quality dimensions in different domains. Knowing the dimensions allows to build a space, but we still need to define a concept. The next step towards that goal would be to define a region in the space. In fact what we are looking for is to define a convex regions (Figure 4.2). Gärdenfors refers to such regions as properties.

4.3.2 Properties and concepts

Properties can go by many different names - features, parameters, qualities, etc. It is not really important how we call it, as long as we know what are we talking about. And to do that lets use the definition that Gärdenfors gives in his theory:

> *A natural property is a convex region of a domain in a conceptual space.*

Even though one might think that since we partition the space into convex regions that we then have clear boundaries. This is not necessarily the case. The definition above does not state that the properties need to have sharp boundaries, but instead the boundaries may by fuzzy, supporting the graded membership idea.

The notion of properties is probably the most understood and used not only in cognitive science but it many other fields as well. As it was pointed out, the similarity is often perceived as simply the sharing of properties, which in the case of media corresponds to sharing content features (for example, sharing certain genres). We use properties all the time in our daily life without even thinking about them, and we can differentiate them even if we do not know the quality dimensions. If we take a color dimensions as an example, we end up with a three dimensional color space. Then an example of a property is a region in that space that we perceive as convex and thus corresponds to a single color - lets say, blue or red.

A property is usually defined with a single dimension, or a small number of integral dimensions. For example a certain weight is an example of a property having only one dimension, while the color red may be understood as a property having three integral dimensions.

The final piece of the puzzle are the concepts. "The standard psychological usage of concept is that of a mental representation individuated or defined by its contents or features" [Laurence and Margolis, 1999]. Gärdenfors defines a concept in the following way:

> *A natural concept is represented as a set of regions in a number of domains together with an assignment of salience weights and information about how the regions in different domains are correlated. [Gärdenfors, 2000]*

A definition of property and concept is quite often confused and other theories sometimes do not make any distinction which is which. In Gärdenfors' theory a property is always based on a single domain, whereas a concept may be (and most often is) based on several different domains. It means that, for example, the properties *red* and *sweet* are based only on single domains – color and taste.

It is important to see the connection between all these different definitions – dimensions, domains, regions corresponding to properties and finally concepts. If we define a concept of an apple as an example then the whole picture would look something like the one in the figure 4.6.

We can actually find examples of properties and concepts when we look at the use of the natural language. We can think of a adjectives as representatives of the properties since they are based on a single domain and aim to identify a certain quality by specifying a certain region in that domain. For example *red*, *sweet*, *round* are all properties (or sticking to the definitions of the chapters 2 and 3, we can call them *features*). Concepts on the other hand are represented by nouns - *a dog*, *a student*, *an apple*. This parallel with the language is helpful, since we all use the nouns and adjectives all the time and we know that an adjective always shows certain features or properties of the object specified by the noun. This is exactly how it works in the conceptual spaces as well.

It is crucial to point out that a concept should not be perceived as just the bag of properties. An important part of knowing the concept means to understand the correlations between the regions from different domains. If we get back to the "apple" example, we can see that there is a strong connection between the region *sweetness* in the domain *taste* and the region *sugar* in the domain *nutrition* [Gärdenfors, 2000].

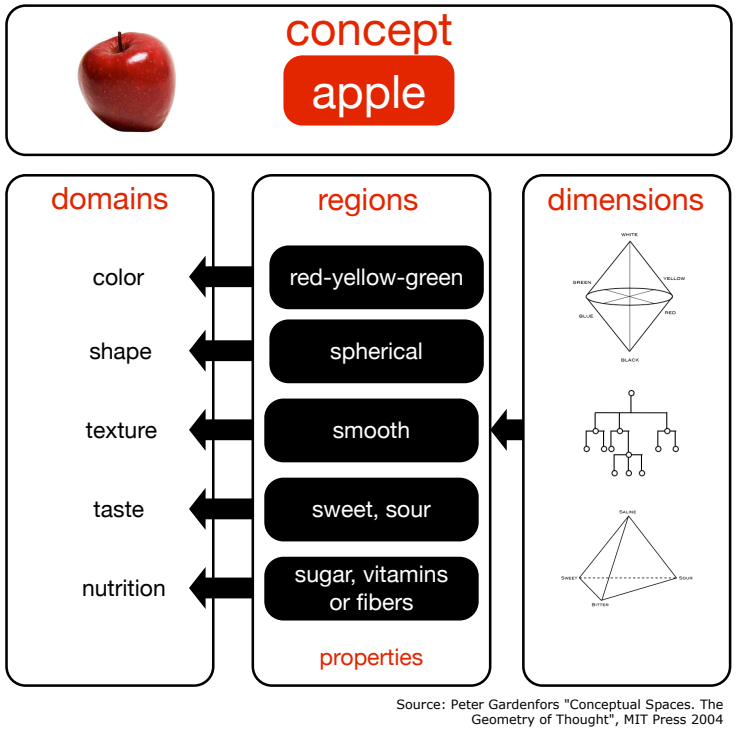


Figure 4.6: An example of the domain regions describing the concept of an apple

Another important point is to say that since properties (adjectives) are describing a concept (a noun), that results in properties being affected by the context, which modifies the concept. It is yet another thing that we take for granted when we speak and understand because our brain in some way is able recognize context and then adjust everything accordingly. For example, we can not talk about a property *tall* or *fast* without specifying the context first. A *tall* regular person is in fact quite *short* basketball player, a *fast* driving a family car is extremely *slow* compared with a Formula-1 race car. We can think of thousands similar examples, because everything is in fact relative. How can we then still talk about properties if it seems that they change their meaning all the time? And if properties are represented as regions in space, does it mean that we need a brand new space for every single concept?

One might argue that properties like the ones in the examples are just too abstract to begin with. How about colors? Can we specify a color (property) without specifying an object (concept) that has that color? Many people would say that we definitely can. If something is identified as being *red* do we then

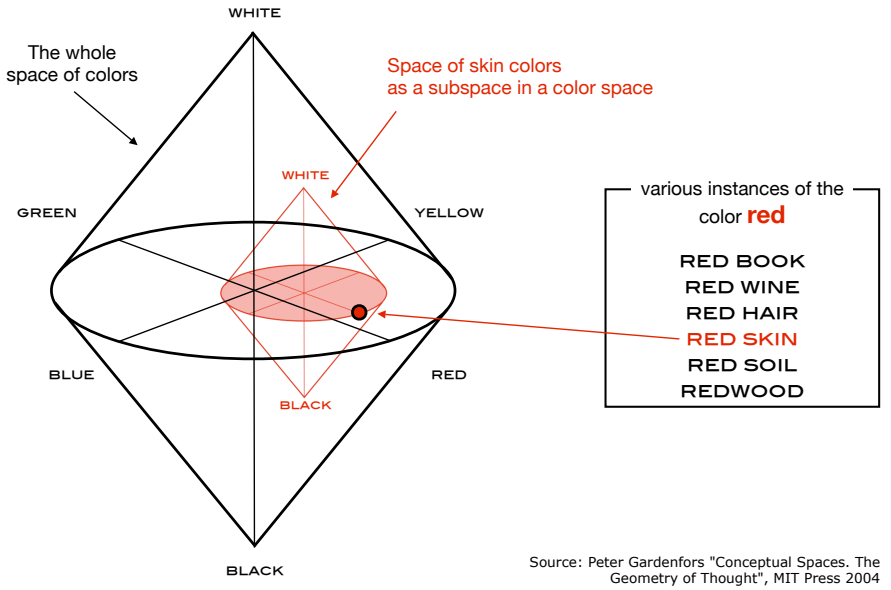


Figure 4.7: An example of the various instances of the color red with one of them (the red skin) is represented as a subspace in a complete color spindle

know exactly what color it is? It may seem that *red* is always *red*, and that there is very little ambiguity here. The truth is a bit more complicated. In the figure 4.7 you see an example of the various instances of the color red [6]. In each of these cases (red hair, red wine, red book, etc...) we use the same word *red* to identify what appear to be quite different colors. How can we we explain this?

In the Conceptual Spaces theory this problem is addressed by using a notion of *contrast classes*, by stating that in terms of cognition what really matters in most cases is the difference - the contrast - between different possible instances rather that their absolute values. While this proposal may seem quite straightforward, the real problem appears when we try to represent what actually happens using the other cognitive theories. Gärdenfors approaches this problem by using subspaces. The main idea is that for every given concept there is a subspace within an original space. The interesting part is that we still keep the same names for the regions based on their relative positions to each other rather that taking their absolute positions in a complete space. Talking about the skin color

[6] Gärdenfors uses this example in his book, but originally it belongs to Clark (1992, pp. 369-372).

[7] people tend to call the lightest region *white* and the darkest *black*, where in fact none of them are even close to the absolute poles of the complete color space. The same goes for other regions as well. In the given example only "red book" uses the whole color space as it is. All the other instances have their own subspaces.

If we get back to media personalization, it was shown in the figure 2.1 that a context information is integral and fundamental part of the personalization system. It was left out of scope in this thesis, but it can be mentioned here that the division of a space into subspace is actually how context could be understood and represented. Of course in order to do that we need to have the spaces, which in most cases we do not. It is an interesting idea nevertheless, and it will probably be a obvious candidate for contextual representation once we manage to to come up with more elaborate and accurate conceptual spaces for media.

Another very important aspect worth mentioning is how the new concepts are learned. "Learning a concept often proceeds by generalizing from a limited number of exemplars of the concept" [Gärdenfors, 2000]. As it was concluded in the previous section, we usually use both exemplars and prototypes for the categorization. Conceptual Spaces theory takes this as a starting point and agreeing that every concept has a prototype which is derived from the exemplars that we encounter. So far it sounds very similar to a Prototype Theory. One of the differences if that Gärdenfors argues that a prototype does not remain rooted at the same spot but it can shift depending on the position of exemplars, compared to a prototype being the same (for example the prototype of a bird being "robin") and forcing exemplars to be categorized based on how close they appear to which prototype.

The figure 4.8 shows an example of how the borders (still keeping the regions convex) of the categories (in this case concepts) are shifting upon an introduction of a new exemplar. In order for this scenario to happen we need to know that exemplar does belong to a a certain concept even if it does appear to be out of boundaries (we need to get feedback). This type of learning is known as *supervised*. The opposite version would be an *unsupervised learning* where no information is given and then everything has to be inferred from analyzing huge data sets of patterns.

[7] To illustrate the subspaces ones needs to have a good idea of what a space looks like to begin with. The problem is that we still do not know many of the dimensions of various spaces, and sometimes we know the dimensions, but are not sure how they fit together. In either case it is an obstacle to actually draw a conceptual space. Color space is relatively simple (low dimensional) and very well understood. That is why color spindle is used to illustrate many of the ideas in the Conceptual Spaces theory.

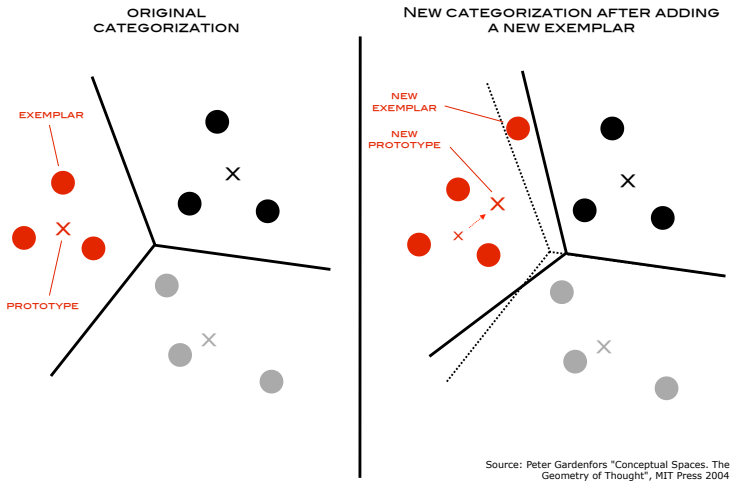


Figure 4.8: An example of concept learning using a Voronoi tessellation – prototype shifts when we introduce a new exemplar into a category.

The final note about this particular knowledge representation theory (application of Conceptual Spaces on emotions to categorize media is one of the main contributions of the thesis) is about where it stands in relation to other cognitive knowledge representation approaches. Talking about the other ones, there are two dominant approaches. First we have the *symbolic* approach which is based on Turing machines where the cognitive process is reduced to calculations and symbol manipulations. The second approach is *associationism*. As an example of this approach we can imagine an artificial neural network where the information is represented in the connections between nodes. Gärdenfors argues that those two approaches should not not be compared directly since they both act on different levels and can be seen a s complementing each other. He puts the Conceptual Spaces theory right in between the two approaches.

4.4 Conclusions

Recommendations are very much related to categorization since in order to recommend we must first know where things belong, and most importantly, where we perceive they belong. This leads to taking a cognitive perspective and trying to get into people's minds to understand what exactly happens when they categorize objects. Three main models of categorization were presented here: classical, prototype and exemplar. The latter two are the more current

and realistic view into how we categorize. The key point from prototypes and exemplars is that categorization, and thus similarity, can be expressed as a distance in a space.

Gärdenfors builds on the these two theories and presents the theory of Conceptual Spaces as cognitive theory of knowledge representation. He talks about concepts and how they are build using properties based on regions in domains having dimensions. In terms of media personalization, his theory suggests that we could express every concept as being positioned in a conceptual space. Since the concepts are based on quality dimensions, that means that if we can express two objects as concepts and map them into space, then we can compare their similarities on a much deeper and more cognitively grounded level.

The next two chapters aim to apply Gärdenfors theory to media, in order to analyze it in terms of its dimensions and knowledge domains, which in return defines properties and concepts.

Chapter 5

Media Personalization Using Genre Metadata

So far a number of different theories and methods were introduced to present a number of different views on the media personalization problem. Looking from the economical point of view the *Long Tail theory* for media has given the motivation and highlighted the need for the personalization, and also has stressed the importance of recommendations as a key component of personalization process. After that the *Chapter 2* presented the state of the art in the area of media recommendation systems. Two main points came out: first, that *similarity* plays the central role in generating recommendations, second, that the additional information about the media (the metadata) seems to be one of the most important areas still leaving a lot of room for improvement. Which naturally led to looking at how do we describe media, and what potential is hidden in different kinds of metadata in terms of similarity knowledge. Up to that point the similarity was simply taken as overlapping features, but is that the only way to look at it? Stepping away from the engineer perspective for a moment, *Chapter 4* has presented that similarity has been a major topic in many philosophical and psychological theories, and therefore introduces a new cognitive perspective on the research question. A cognitive theory of Conceptual Spaces [Gärdenfors, 2000], which can be seen as a theory of knowledge representation, is one of a few where the notion of similarity is combined with the actual representation of object by using conceptual spaces. This is more or less as far as I have gone theoretically. Now it is time to combine all these different

elements and apply them in a practical way.

The following two chapters present three different cases. It all begins in the video domain (TV to be precise) by using the *genre* information which represents the traditional structured approach. After that, in *Chapter 6*, I shift to a very unstructured type of metadata – *synopsis* which can be seen as being "closer" to the content and focuses on describing the meaning rather than trying to categorize it. Then I take another step further and cross the boundary between metadata and the actual content thus getting as close as we can to media. Song *lyrics* are a perfect example of such "metadata/content", therefore here I turn into the music domain – the second part of the *Chapter 6*. All along the way I stick to the core idea of similarity as a basis for personalization and try to apply Conceptual Spaces for each of the three cases. This is the very top level outline of where the rest of the thesis goes. Now lets start from the beginning.

One final note – most of the results presented in the chapters 5 and 6 builds on the authors earlier and current research done in collaboration with Michael Kai Petersen (DTU–IMM). Therefore in a number of places I refer to the authors as *we*.

5.1 TV-Anytime Genre Metadata

As it was introduced in the chapter 3, TV-Anytime (TVA) is a metadata specification originally created for describing TV content. With its phase 2 release TVA has expanded its focus and now it is not limited to a TV domain, but instead serves as a universal audiovisual media description specification. I selected TVA a a part of the empirical data for a few reasons. First of all, TVA represents the state of the art in terms of media description in a number of aspects. It contains all different kinds of metadata – editorial, technical and content, the latter being quite elaborate compared with other alternatives on the market. TVA content-related metadata includes all the creation related information (titles, names, etc.), plus it supports synopsis information (which is quite common in TV world), keywords and on top of that it has its own [1] Classification Schemes (CS) in order to divide the content into predefined categories. TVA has 16 of such classification schemes [2] each taking a different angle of how we can look at a TV program.

[1] Even though they are different from MPEG-7 Classification Schemes, they are still heavily influenced by MPEG-7.

[2] Phase 2 added a number of new Classification Schemes, but none of them are related to the actual content, but instead focus on different contextual information this expanding the scope of TVA into other domains, such as games.

The second reason of choosing TVA was triggered by the fact that there is plenty of empirical data available since the biggest promoter of TVA – BBC – has made all its content metadata freely available and for the last 3 years has been putting on *backstage.bbc.co.uk* website.

Without looking at the synopsis information, classification schemes are the most descriptive type of metadata that TVA has to offer. Sometimes classification schemes will be simply referred as *Genre* metadata, because in TVA classification schemes are used within the <Genre> element. Therefore TVA *Genre* is not what we traditionally think and use as genres in music or movies, although there is an overlap. Before trying to build a concept using genres lets take a look at what kind of data are we actually dealing with.

	LEVELS			
TVA CS	1	2	3	SUM
CONTENT	6	22	59	87
INTENTION	4	7		11
FORMAT	5	9	2	16
INTENDED AUDIENCE	1	4	2	7
ATMOSPHERE	34			34
SUM	50	42	63	155

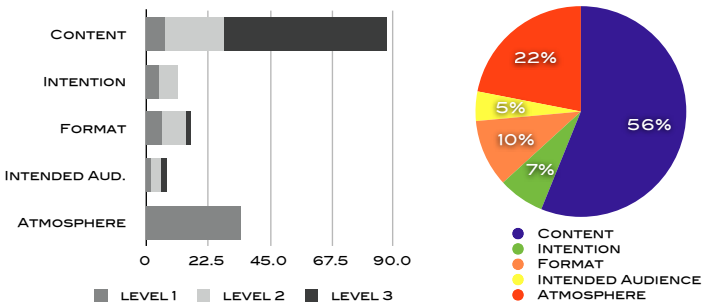


Figure 5.1: Data showing how many genres of each CS is BBC using.

From all of the Classification Schemes there are five that are targeted purely to describe content and therefore it is no surprise that in the BBC implementation of TVA, they use only these five CS: *Content*, *Intention*, *Format*, *Intended Audience* and *Atmosphere*. To give an idea of how big these schemes are, it is interesting to notice that all five combined together have 922 unique genres spanned over a up to four-level hierarchies. BBC uses roughly only one sixth of all the genres in those five CS (see Figure 5.1). Since all the empirical data in this part is taken from the BBC, from now on when talking about TVA I will only refer to the BBC subset of TVA.

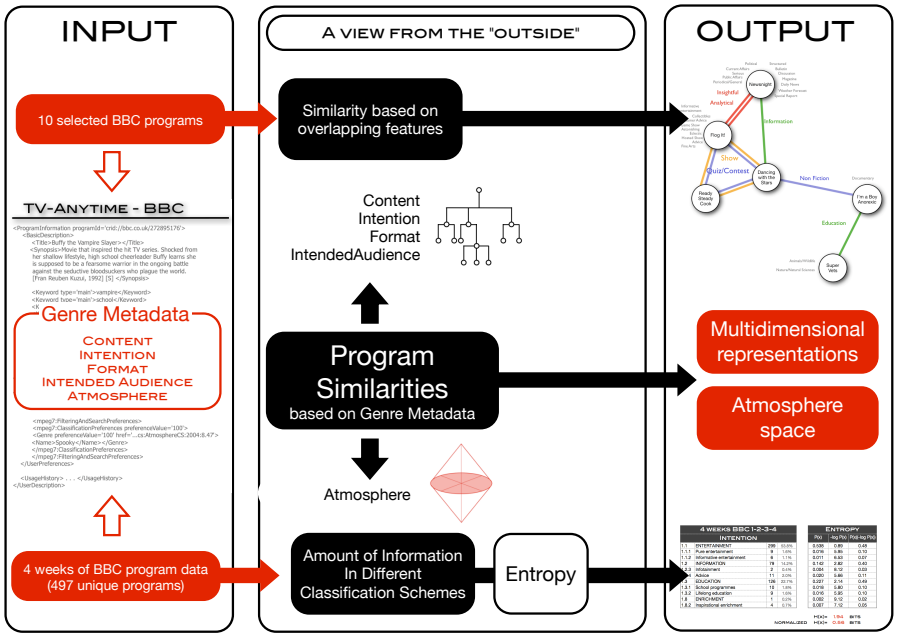


Figure 5.2: An overview of the methodology to analyze TV programs based on their genre metadata

It can be seen that *Content* has the widest range of genres practically used, with most of them located three levels deep. This only shows the variety of selection and does not reflect how often certain genres are used. Imagine that every program is annotated only with a set of genres from various CS. Then the most similar programs are the ones that share many of their genres. Here genres act like features used to describe a program, or drawing a reference to the Conceptual Spaces – properties. In order to analyze the TV programs using genres the following methodology is proposed (see Figure 5.2).

To see the big picture of how genres are actually used, the empirical data – all the TV programs broadcasted in the period of one month for four channels – was taken from the BBC. That corresponds to nearly 497 TV programs for four of the main channels: BBC One, BBC Two, BBC Three and BBC Four. Data itself is serialized in XML as a common standard for this type of purpose, therefore in order to extract the actual genres a number of XSLT stylesheets were built. Two data sets were used - one consisting of one month worth of programs to show general tendencies, another consisting of only 10 preselected programs in order to be able to focus on program overlaps. One of the key questions here is to evaluate how much similarity knowledge we have in genres

and whether that results in the meaningful overlaps. Eventually it takes us to the question – can we build a conceptual space using genres, or maybe only using genres from certain Classification Schemes.

5.2 The structure and topology of TVA genres

The one classification scheme that is found is nearly every program is *Content*. This scheme shows the type of the program and represents what people traditionally think when talking about a TV genre. As can be seen from the genre distribution (see Figure 5.3) there are a few genres that are used very often and there are many of them that are used only a few times. This illustrates a very typical problem found in many detailed classification schemes. If we are bound by retrieving programs using overlapping genres then we are limited to a very high level categories that do not say much about the content. On the other hand, much narrower and thus more informative genres would be a bit more helpfull but they simply occur too few times to generate any significant overlaps. For example genres *Non-Fiction/Information*, *Fiction/Drama* and *Amusement/Entertainment* are found very often, but because being too general they do not provide much help, whereas genres like *Spiritual* or *Cultural* are found very rarely.

A measure of entropy is used here to show the average amount of information that a certain Classification Schemes has based on the probabilities of its genres. The higher the entropy – the more unpredictable the outcome (the genre) of the source (the Classification Scheme) is, meaning that we get more information if we are told that a program has genre *Spiritual* compared to *Fiction/Drama* since the earlier one is less probable is results in the higher surprisal. Since every CS has a different number of genres, in order to be able to compare them we need to normalize the entropy by dividing it by $log_2(n)$, where n is the total number of different genres.

the total number of programs	497	
the number of programs that have this CS	485	98%
the total number of genres used	894	
the number of unique genres	64	
the average number of genres per program	1.8	
the average no. of programs connected by one genre	14.0	

4 WEEKS BBC 1-2-3-4 CONTENT				ENTROPY		
				P(x)	-log P(x)	P(x)(-log P(x))
3.1	NON -FICTION/INFORMATION	137	15.3%	0.153	2.71	0.41
3.1.1.1	Daily news	8	0.9%	0.009	6.80	0.06
3.1.1.3	Special report	1	0.1%	0.001	9.80	0.01
3.1.1.5	Periodical General	15	1.7%	0.017	5.90	0.10
3.1.1.7	Economy/Market/Financial	1	0.1%	0.001	9.80	0.01
3.1.1.9	Sport News	6	0.7%	0.007	7.22	0.05
3.1.1.10	Cultural	5	0.6%	0.006	7.48	0.04
3.1.1.11	Local/regional	1	0.1%	0.001	9.80	0.01
3.1.1.12	Traffic	1	0.1%	0.001	9.80	0.01
3.1.1.13	Weather forecasts	8	0.9%	0.009	6.80	0.06
3.1.1.15	Public affairs	2	0.2%	0.002	8.80	0.02
3.1.1.16	Current affairs	4	0.4%	0.004	7.80	0.03
3.1.2	Philosophies of life	4	0.4%	0.004	7.80	0.03
3.1.2.1	Religion	1	0.1%	0.001	9.80	0.01
3.1.3.1	Political	10	1.1%	0.011	6.48	0.07
3.1.3.5	Finance	1	0.1%	0.001	9.80	0.01
3.1.3.6	Education	10	1.1%	0.011	6.48	0.07
3.1.4	Arts and Media	26	2.9%	0.029	5.10	0.15
3.1.4.5	Cinema	1	0.1%	0.001	9.80	0.01
3.1.4.9	Fine Arts	2	0.2%	0.002	8.80	0.02
3.1.5.3	History	12	1.3%	0.013	6.22	0.08
3.1.6	Sciences	8	0.9%	0.009	6.80	0.06
3.1.6.2	Nature/Natural sciences	20	2.2%	0.022	5.48	0.12
3.1.6.3	Animals/Wildlife	20	2.2%	0.022	5.48	0.12
3.1.6.4	Environment/Geography	3	0.3%	0.003	8.22	0.03
3.1.6.8	Technology	8	0.9%	0.009	6.80	0.06
3.1.6.12	Spiritual	1	0.1%	0.001	9.80	0.01
3.1.7.7	Religious buildings	1	0.1%	0.001	9.80	0.01
3.1.9.2	Fair	1	0.1%	0.001	9.80	0.01
3.2	SPORTS	31	3.5%	0.035	4.85	0.17
3.3	LEISURE/HOBBY	1	0.1%	0.001	9.80	0.01
3.3.1	Do-It-Yourself	1	0.1%	0.001	9.80	0.01
3.3.3	Gardening	1	0.1%	0.001	9.80	0.01
3.3.7	Outdoor	1	0.1%	0.001	9.80	0.01
3.3.20	Life/House Keeping/Lifestyle	1	0.1%	0.001	9.80	0.01
3.3.26	Consumer advice	3	0.3%	0.003	8.22	0.03
3.3.31	Collectibles	2	0.2%	0.002	8.80	0.02
3.4	FICTION/DRAMA	102	11.4%	0.114	3.13	0.36
3.4.1	General light drama	59	6.6%	0.066	3.92	0.26
3.4.2	Soap	2	0.2%	0.002	8.80	0.02
3.4.2.1	Soap opera	3	0.3%	0.003	8.22	0.03
3.4.5	Medical melodrama	3	0.3%	0.003	8.22	0.03
3.4.6.4	Detective	1	0.1%	0.001	9.80	0.01
3.4.6.7	Science fiction	7	0.8%	0.008	7.00	0.05
3.4.15	Contemporary drama	2	0.2%	0.002	8.80	0.02
3.5	AMUSEMENT/ENTERTAINMENT	151	16.9%	0.169	2.57	0.43
3.5.1	Game Show	4	0.4%	0.004	7.80	0.03
3.5.2	Quiz/Contest	24	2.7%	0.027	5.22	0.14
3.5.7	Comedy	81	9.1%	0.091	3.46	0.31
3.5.7.3	Sitcom	4	0.4%	0.004	7.80	0.03
3.5.7.4	Satire	1	0.1%	0.001	9.80	0.01
3.5.9	Humour	5	0.6%	0.006	7.48	0.04
3.6	MUSIC	6	0.7%	0.007	7.22	0.05
3.6.1	Classical music	5	0.6%	0.006	7.48	0.04
3.6.1.14	Choral	1	0.1%	0.001	9.80	0.01
3.6.2	Jazz	1	0.1%	0.001	9.80	0.01
3.6.4.1	Pop	3	0.3%	0.003	8.22	0.03
3.6.4.4	Soft/Slow Rock	2	0.2%	0.002	8.80	0.02
3.6.4.5	Classic/Dance/Pop-rock	7	0.8%	0.008	7.00	0.05
3.6.6	Country and western	2	0.2%	0.002	8.80	0.02
3.6.8	Electronic/Club	1	0.1%	0.001	9.80	0.01
3.6.12.3	Ceremonial/Chants	1	0.1%	0.001	9.80	0.01
3.6.13	Spoken	1	0.1%	0.001	9.80	0.01
3.8	EDUCATIONAL MATERIAL	56	6.3%	0.063	4.00	0.25

H(X)= 4.29 BITS

NORMALIZED H(X)= 0.72 BITS

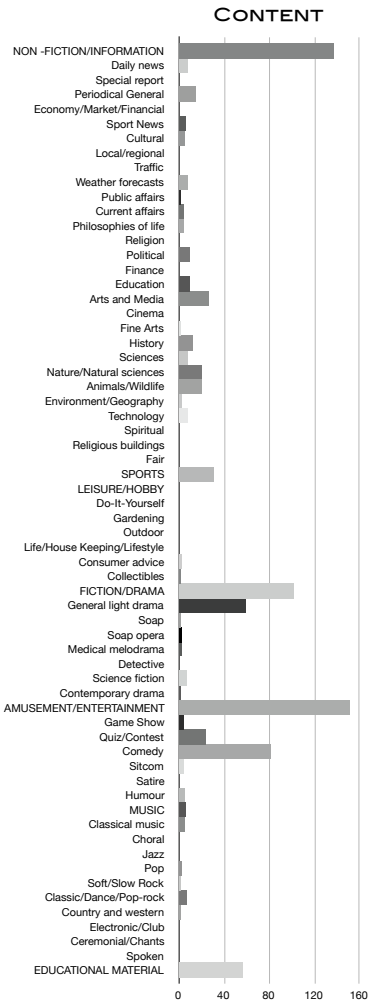


Figure 5.3: A distribution of genres according to their usage in the Content CS with the average level of information that this taxonomy holds

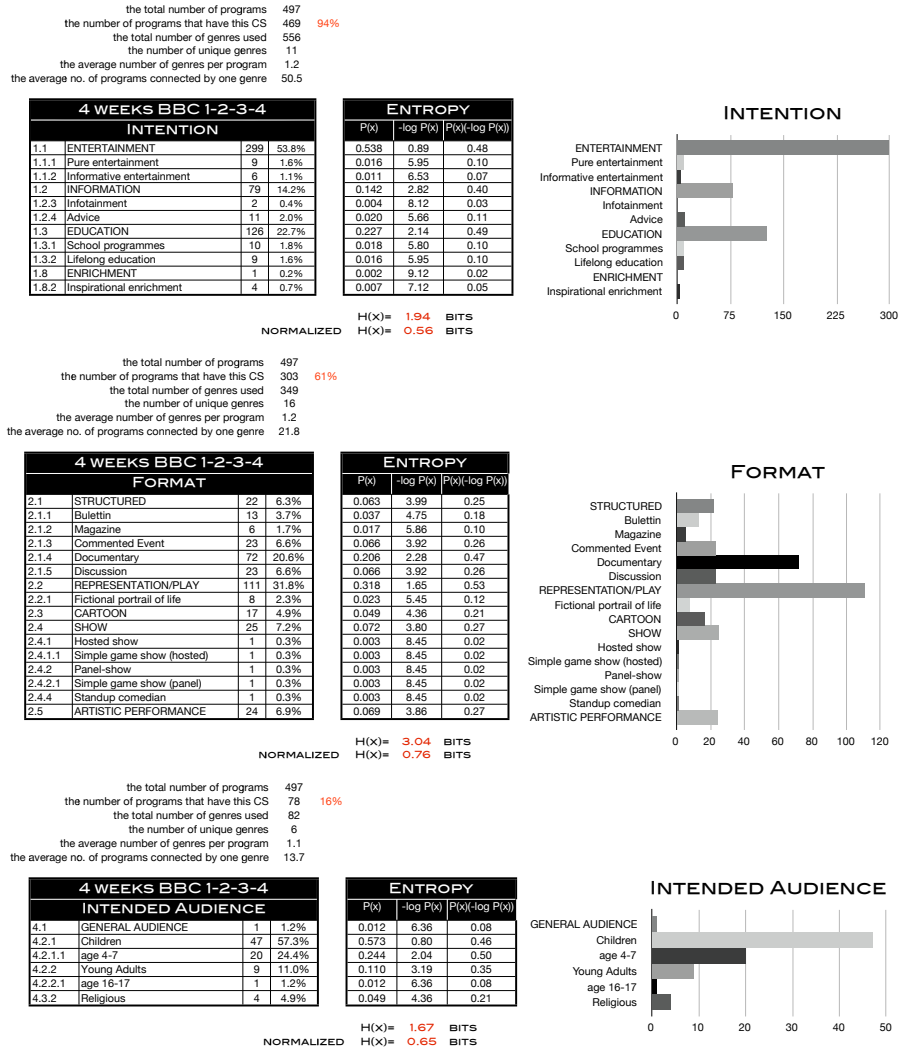


Figure 5.4: A distribution of genres according to their usage in the Intention, Format and Intended Audience CS with the average level of information that those taxonomies hold.

the total number of programs	497	
the number of programs that have this CS	28	6%
the total number of genres used	81	
the number of unique genres	31	
the average number of genres per program	2.9	
the average no. of programs connected by one genre	2.6	

4 WEEKS BBC 1-2-3-4 ATMOSPHERE			
8.1	ALTERNATIVE	2	2.5%
8.2	ANALYTICAL	8	9.9%
8.3	ASTONISHING	3	3.7%
8.6	BREATHTAKING	1	1.2%
8.7	CHILLING	1	1.2%
8.1	CONFRONTATIONAL	1	1.2%
8.11	CONTEMPORARY	5	6.2%
8.12	CRAZY	1	1.2%
8.14	ECLECTIC	2	2.5%
8.16	EXCITING	1	1.2%
8.19	FUN	3	3.7%
8.2	GRIPPING	2	2.5%
8.21	GRITTY	4	4.9%
8.22	GUTSY	3	3.7%
8.24	HEART-RENDING	1	1.2%
8.25	HEART-WARMING	6	7.4%
8.27	HUMOROUS	5	6.2%
8.29	INSIGHTFUL	10	12.3%
8.31	INTRIGUING	1	1.2%
8.32	IRREVERENT	3	3.7%
8.33	LAID BACK	1	1.2%
8.35	PEACEFUL	1	1.2%
8.37	PRACTICAL	3	3.7%
8.38	ROLLERCOASTER	1	1.2%
8.4	ROUSING	1	1.2%
8.42	SATIRICAL	1	1.2%
8.43	SERIOUS	1	1.2%
8.46	SILLY	3	3.7%
8.47	SPOOKY	2	2.5%
8.48	STUNNING	1	1.2%
8.53	WACKY	3	3.7%

ENTROPY		
P(x)	-log P(x)	P(x)(-log P(x))
0.025	5.34	0.13
0.099	3.34	0.33
0.037	4.75	0.18
0.012	6.34	0.08
0.012	6.34	0.08
0.012	6.34	0.08
0.062	4.02	0.25
0.012	6.34	0.08
0.025	5.34	0.13
0.012	6.34	0.08
0.037	4.75	0.18
0.025	5.34	0.13
0.049	4.34	0.21
0.037	4.75	0.18
0.012	6.34	0.08
0.074	3.75	0.28
0.062	4.02	0.25
0.123	3.02	0.37
0.012	6.34	0.08
0.037	4.75	0.18
0.012	6.34	0.08
0.012	6.34	0.08
0.037	4.75	0.18
0.012	6.34	0.08
0.012	6.34	0.08
0.012	6.34	0.08
0.012	6.34	0.08
0.012	6.34	0.08
0.037	4.75	0.18
0.025	5.34	0.13
0.012	6.34	0.08
0.037	4.75	0.18

H(x)= 4.55 BITS
NORMALIZED H(x)= 0.92 BITS

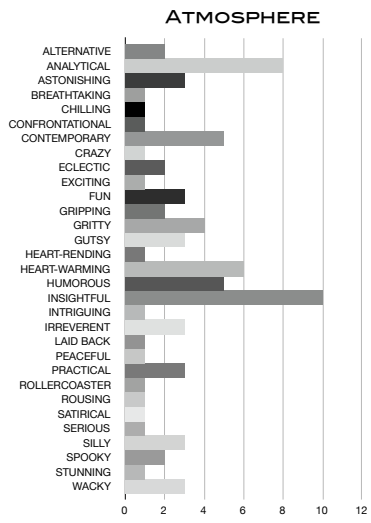


Figure 5.5: A distribution of genres according to their usage in the Atmosphere CS with the average level of information that this taxonomy holds

Another three classification schemes – Intention, Format and Intended Audience – try to describe another aspects that a TV program has, such as what is the intention of the program, what is the format or which audience is it targeted to. Their usage distribution is similar to the one of the *Content*, since they each have a few dominant genres that tend to be used over and over again (see Figure 5.4). *Intention* turns out to be the least informative CS having the smallest normalized entropy (0.56 bits) since it has three very dominant genres: *Entertainment*, *Entertainment* and *Entertainment*. No surprise that all of them are on the first level of the hierarchy. Given the fact that every program has on average 1.2 *Intention* genres, one could expect to see lots of programs overlapping on these three genres and nothing much on the lower levels.

The fifth Classification Scheme is the *Atmosphere* and it aims to describe the emotional response of the program rather than putting it into a certain category. It can be seen that this scheme has is used much less than the other ones even though it has relatively high number of genres that have been used at least once (see Figure 5.5). *Atmosphere* genre usage is much more uniform, which could be partially explained by the fact that this particular scheme has only one level with all the genres presented as a bag of features instead of having a

tree topology like the other CS. This leads to this scheme having the highest normalized entropy compared to the other ones. As it was discussed in chapter 3, such emotional metadata can be very useful because it reflects the inner qualities of the media that people perceive as very important. We will get back to this discussion while evaluating the actual program overlaps.

There is some some extra information given about each CS right above each of the the tables (the six lines). The most interesting are the two lowest lines *the average number of genres per program* and *the average number of programs connected by one genre*. Although both of these numbers show the averages and do not reflect individual genres, they provide general information about what can we expect from a certain CS. For example it can be seen that 98% of all analyzed TV programs (497 programs) have *Content CS* and that on average one program is described by nearly two *Content* genres. 94% of programs also have the *Intention CS*, 61% have *Format CS* and only 16% have *Intended Audience CS*. Interesting thing to notice is that the programs that have genres from any of these three Classification Schemes (*Intention*, *Format* and *Intended Audience*) they have on average only one genre from each CS. This shows that in most cases it is perceived to be enough to assign only one genre from each CS to a program since in most cases the genres of the same CS do not overlap.

Atmosphere CS is a bit different. Only 6% of all programs use it, but those that do, have on average 3 different *Atmosphere* genres assigned to them. This can be partially explained by the fact that TVA treats *Atmosphere* as a bag of features with no hierarchy and that this leads to people using more terms in parallel whereas all other CS have a tree structure which suggest that quite often people would simply put a program on a single branch.

5.3 TV program similarity based on genres

Now that we know the general tendency of genre distribution, lets take a look at how much programs overlap when using these genres. For this purpose a much smaller set of programs were taken selecting a sample of programs where some of them have lots of genre metadata (for example *Flog It!*, *Newsnight*, *Little Britain* and *Two Pints of Lager And*) and also throwing a few programs into the mix that would could be perceived as similar but are not that heavily annotated (for example *The Flying Gardener*). On top of that those programs were required to have a synopsis because I wanted to use them again in the second part and see if we can get more information this way. As a result 10

programs were selected (see Figure 5.6)[3].

Since TVA does not specify any relationships between different CS, I will present the 10 program overlaps for every CS one at a time. And after that I will try to sum everything up. None of these programs had any *Intended Audience* genres therefore the overlaps are presented only for the remaining four CS.

	Buffy	Dancing With Stars	Super Vets	The Flying Gardener	Flog It!	NewsNight	Ready Steady Cook	I'm A Boy Anorexic	Little Britain	Two Pints of Lager
Buffy		1			1		1		2	1
Dancing With Stars	1				4	1	4	1	2	2
Super Vets								1		
The Flying Gardener										
Flog It!	1	4				2	4		2	2
NewsNight		1			2					
Ready Steady Cook	1	4			4				2	2
I'm A Boy Anorexic		1	1							
Little Britain	2	2			2		2			8
Two Pints of Lager	1	2			2		2		8	
has genres	8	6	3	1	15	15	4	3	12	14
overlaped programs	5	7	1	0	6	2	5	2	5	5
no. of genres used	2	6	1	0	6	3	4	2	9	8
% of genres used	25%	100%	33%	0%	40%	20%	100%	67%	75%	57%

Figure 5.6: 10 selected programs with numbers identifying how many genres are shared between each of the program pairs.

In order to present the program overlaps I chose the model that reminds of the different energy levels in the atom. If we follow this allegory, then in the center of each diagram we have a Classification Scheme acting as a nucleus, whereas all the 10 programs are located around it like electrons [4]. Only genres of that particular CS are shown in each of the diagrams, where the genres that are shared are shown in respective color and genres that are not shared are just shown in small font in black. The size of the font of a shared genre signified by how many programs it is shared. Programs are grouped together united by shared genres. The actual location of the program does not matter, what matters are the two things. First, the links between programs identified by transparent clouds. Second, the distance between the program and the center shows how many "active" genres of that particular CS it shares with other programs. Programs outside the diagrams are the ones that happened not to share any of their genres with others. And finally, if the program does not even have a single genre in a particular CS, it is still shown outside the circle, only in gray color.

[3]The table is symmetric, therefore it could have been presented using only half of it, but I chose to show it like this since I wanted to present several statistical measures at the bottom of it.

[4]Even though this thesis is using many different theories, nuclear physics is not one of them. It just seemed a clear analogy and I took a risk to use it without any intent to confuse anybody (no, the programs do not circle around the center)

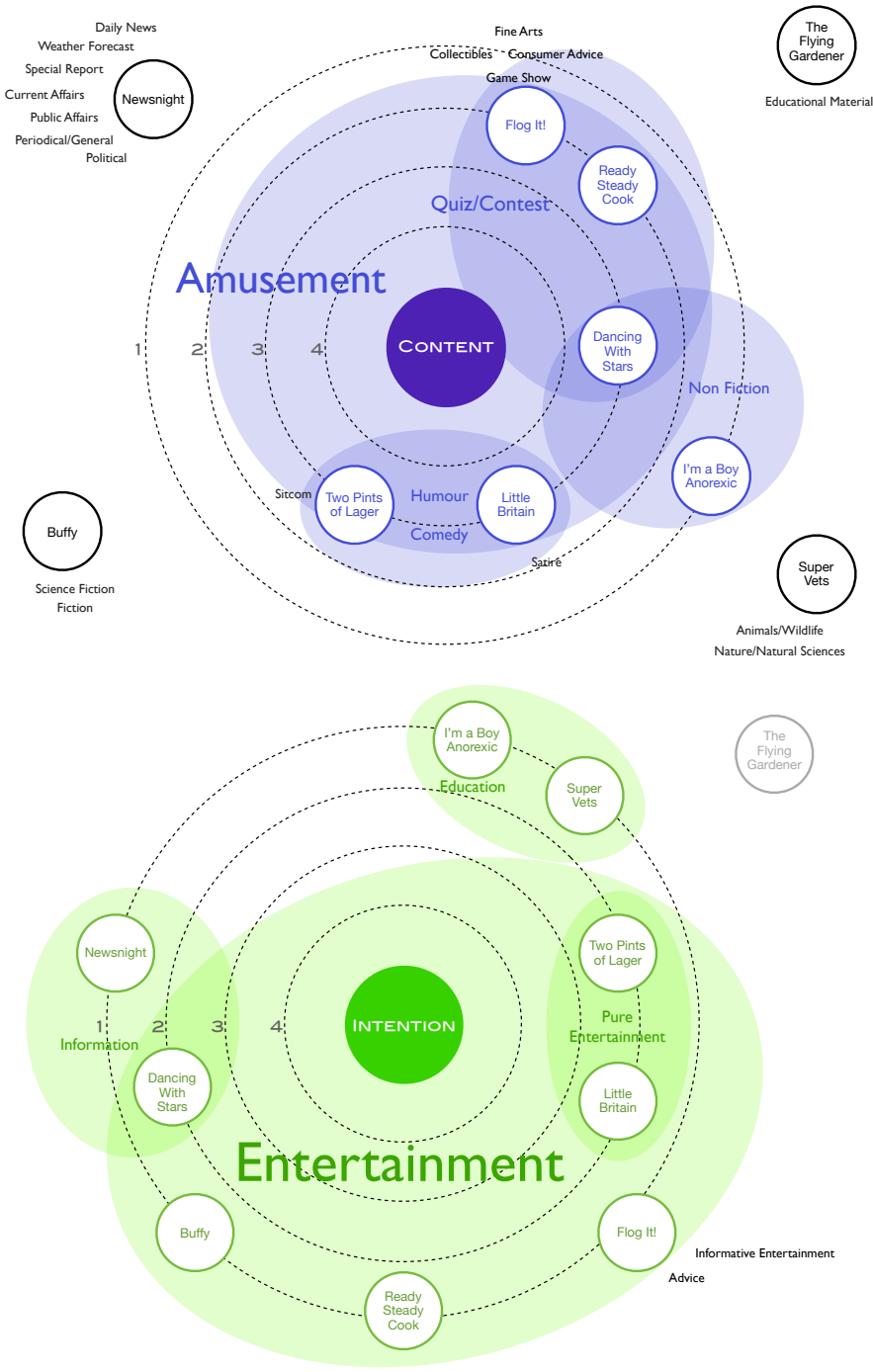


Figure 5.7: Program overlaps for *Content* and *Intention* CS

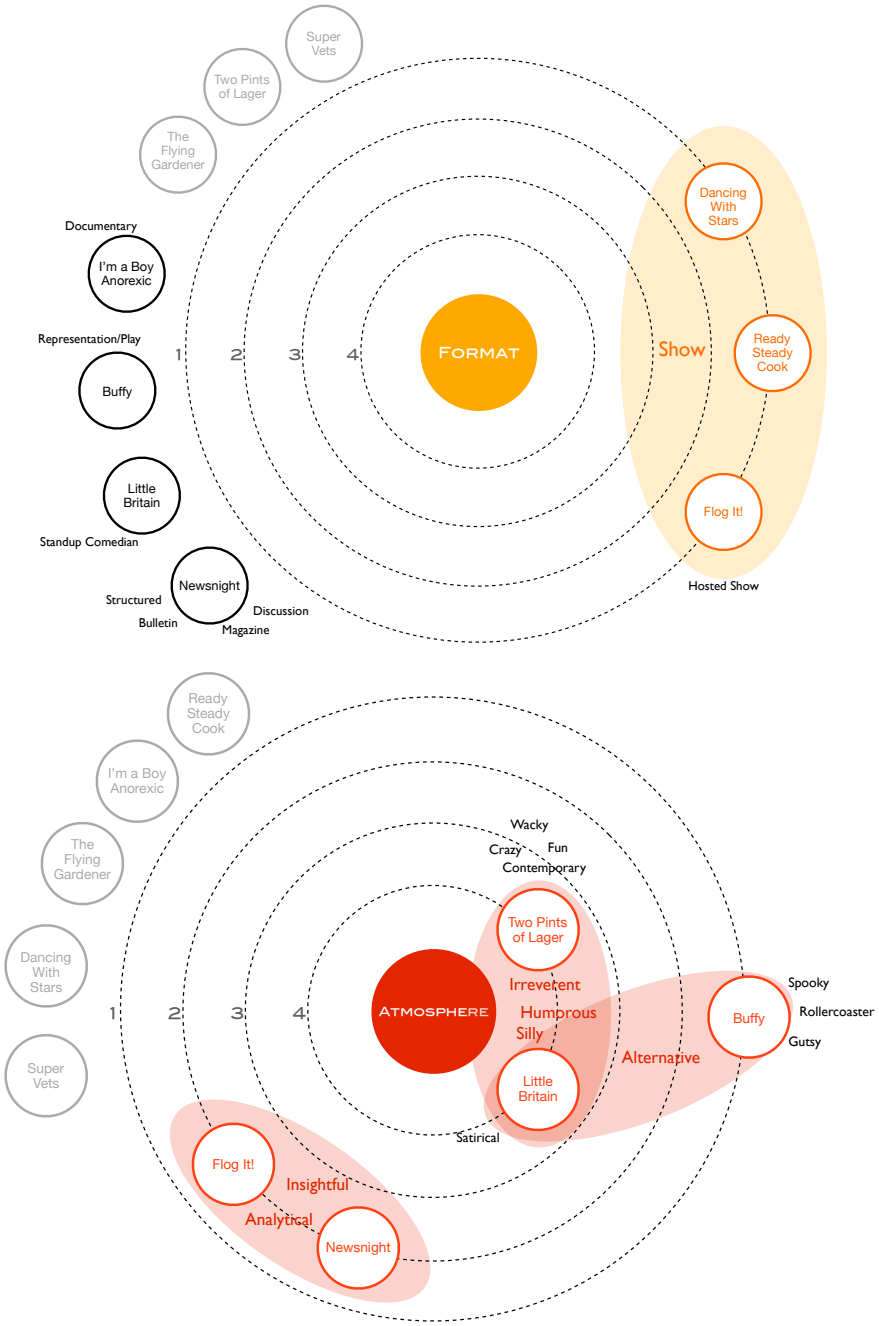


Figure 5.8: Program overlaps for *Format* and *Atmosphere* CS

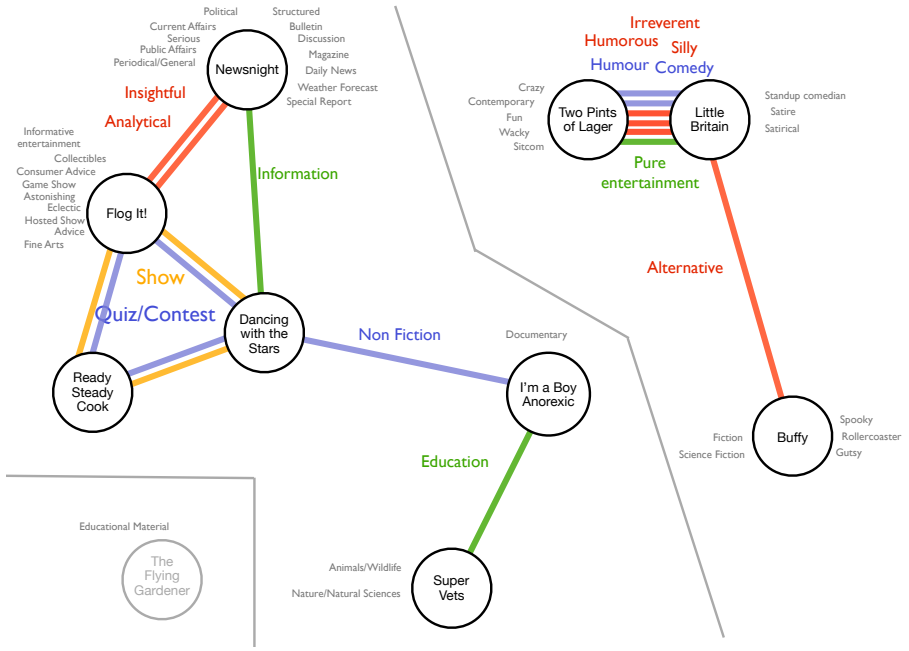


Figure 5.9: Connections among programs based only on them sharing genres in *Content*, *Intention*, *Format* and *Atmosphere* CS.

What we see from these four diagrams only confirms the assumptions drawn from looking at the general genre distributions earlier. We see huge overlaps on very general genres like *Entertainment* or *Amusement* (see Figure 5.7)(one could argue that such genres do not carry much, if any, meaning since they are very vague). Next to that we see a number of more precise genres that are not shared, even though the programs seem to have similarities. Of course, this is a very small set and one should be very careful drawing any conclusions from it. But if we look in the earlier bigger data set, we can see the exact same problems.

Another interesting observation is that the CS with a tree topology – *Content*, *Intention*, *Format* (I asume this would be the same with *Intended Audience* as well) – seem to offer programs that stay in the same domain and lead to having "more of the same" type of recommendation. Whereas the *Atmosphere* seems to span across domains rather that staying in them (see Figure 5.8). One solution is to add more metadata descriptions in parallel belonging to different domains but whether this facilitates identifying similar programs depends on these dimensions being orthogonal so that they would add new information rather than stating what is already known. This idea is discussed further in the first publication that is included in the thesis on this particular case (see

Appendix: A).

Lets put all of these different CS together and see how close the programs get to each other judging on mere genre overlap (see Figure 5.9).

Programs *Flog It!* and *Newsnight* are an example of how two programs that each have way over an average amount of genres in different CS are overlapped only by an *Atmosphere* because they seem to share an overall approach and mood rather than content. This again makes *Atmosphere* a very special Classification Scheme. At this point in the thesis we know that emotional metadata is very important and people use it quite a lot when they are given a chance to annotate content themselves (in folksonomies).

5.4 Building a concept of a TV program using genres

Data suggests that there are severe limitations in terms of finding similar TV programs if we only use genres and nothing else. In most cases this is exactly what is used. As an example one could think of an Electronic Program Guide (EPG) on your TV screen where you can browse programs by genres, or when you can specify genres in your TV-Anytime User Profile. None of these situations give the wanted results, which would be – to find similar yet novel programs. Basing similarity on genres alone seems similar to the classical Aristotelian categorization method. But now we know that there are other alternatives, graded memberships and family resemblances, we have spaces where we can put objects and then use the distance as a measure of similarity. Lets try to apply the Conceptual Space to a TV program and see if it makes any sense. I will use the same template as I did to illustrate the concept of an apple in the previous chapter (see Figure 4.6), so that it would show TVA metadata fits where according to the Conceptual Spaces theory.

The Conceptual Spaces theory states that a concept representation first of all must have a set of regions in a number of domains. This is very much like it is in TVA Classification Schemes where regions would refer to individual genres, while the domains would be expressed as the Classification Schemes themselves. But domains must be formed by using integral dimensions. Is that the case in TVA? To some extent, yes. If we assume that the entire CS is based on a single dimension, then it can definitely form a valid region. For example, *Content* can be understood as a single dimensional tree, where all the nodes are regions on the same dimension. Since *Intention*, *Format* and *Intended Audience* all are represented as a single tree, then we can assume that there is one dimension

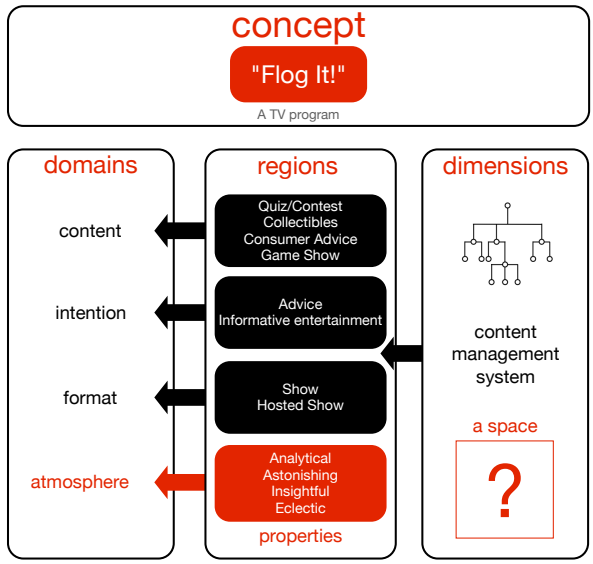


Figure 5.10: The conceptual representation of a program *Flog It!*.

and that we do not need to check the integrity since there is no other dimension there (see Figure 5.10).

But we could also look at it from another angle. All these trees do not have a single root element. It may be assumed that a root element is the Classification Scheme itself, but that still is only the assumption with no cognitive grounding. In some cases it makes more sense to think of certain Classification Schemes as a collection of dimensions, each represented in a tree, or some other topology (linear, binary). Imagine every first level genre being a root element for all the other descendant genres. If that is the case then we must insist that dimensions have to be integral and if they are not then we must divide a Classification Scheme in such a way that it would be either a single dimension, or that dimensions would be integral. For example the *Intended Audience* CS contains information about gender, occupation, age. All these are valid dimensions, but they are definitely not integral. That would mean that we can not represent *Intended Audience* in one coherent space. Other Classification Schemes present their own challenges since a good understanding of which dimensions are used and how they look, and most importantly how they interact, is essential for concept formation. Here I can only agree with Gärdenfors and say that we do need to gather more knowledge about dimensions and how they look in the case of current TVA Classification Schemes.

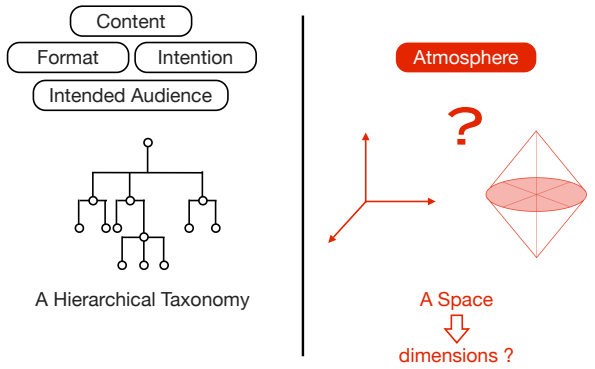


Figure 5.11: Division of TVA CS according to their topologies.

So far we saw that all the CS are based on a tree structure that may or may not be cognitively grounded. All except *Atmosphere*. In TVA *Atmosphere* is simply represented on one level without any structure (see Figure 5.11). This should not be understood the way that these terms are all equally related. I assume that people in TVA either did not know about the possible ways to represent emotions, or they figured out that since the genres are assigned manually then it would be much easier to simply give a list of emotional terms and let people pick from it, rather than engaging in a much more complex procedure as building a space. Lets take a closer look into the *Atmosphere* CS and see if we can see some dimensions there.

When *Atmosphere* genres are presented in such a way (see Figure 5.12), we immediately start seeing some structure and that suggest that there must be some kind of dimensions there. Here genres, representing the emotional terms, are grouped both vertically and horizontally. Vertical axis show the level of user involvement, going from passive "couch potato", all the way to very engaged and ready to interact. Horizontally genres are grouped according to their similarities, representing the intensity of the emotion. For example starting with *humorous* and ending up with *crazy*. There are also contrasts, for example *heart-rending* versus *heart warming*. This is only my interpretation of how these genres could fit together. At this point this is neither valid, nor grounded cognitively. The goal here is simply to show that there is structure when we talk about emotions, and if the dimensions turn out to be integral then we can talk about building an emotional space.

As I mentioned in chapter 3, all the TVA genres could be extracted from a content management system, since this is a way to automate the process of annotation, at least partially. Normally we will not find emotional terms in the

Figure 5.12: Atmosphere genres presented as being grouped according to their similarities or contrasts

content management system. Then the only way is to have somebody to sit down and annotate manually. This solution sounds neither too good nor too accurate. An alternative would be to extract emotional metadata automatically. And if there is one place where TV programs have it, it is the *Synopsis*. Which brings us to the next section, where I will try to see if we can improve the concepts formation procedure by using emotions extracted from the synopsis instead of genres.

5.5 Conclusions

This chapter has demonstrated that using only TVA *Content* Classification Scheme as a basis for categorization of programs will tend to identify programs which belong to closely related categories leading us to "more of the same" instead of something "similar" in its essence rather than a content. The same goes for *Intention*, *Format* and *Intended Audience* CS. Programs which share genres belonging to the very top level of a content taxonomy provide a too general categorization in order to identify similar programs, while adding more details from the lower layers will only decrease the amount of shared features and thus fail to recommend any programs at all. The focus here, much like

in the whole thesis is not on media search, but on media recommendations. Therefore the main goal is not to find precisely what one is looking for, but to be able to automatically recommend similar items.

A possible solution might be to combine metadata descriptions in parallel belonging to different domains but whether this facilitates identifying similar programs depends on these CS being orthogonal so that every CS would add information rather than just stating the same thing that has been already stated by another CS.

It was discussed whether these genres (properties) when combined could be organized into a conceptual space. It turns out that the TVA genres may not be necessarily based on integral dimensions therefore limiting the very formation of the concept.

The one CS that is different in a number of ways seems to be interesting and promising in terms of the program similarities that it can provide. Such CS is the emotional terms gathered under the *Atmosphere* CS. It was suggested that *Atmosphere* could be organized in a space rather than being without structure the way TVA sees it.

Chapter 6

Media Personalization Using Affective Terms

The previous chapter left us with the idea that it may be possible to build conceptual spaces for media using metadata. The main problem turned out to be the quality dimensions of media classification schemes. It is perfectly understandable because TVA Classification Schemes are an example of a structured metadata that was not built according to cognitive theories but instead it followed the traditional librarian approach where everything is organized into the neat little categories forming huge and complex structures that are not necessarily cognitively grounded. The ray of light turned out to be the very unstructured *Atmosphere* Classification Scheme which uses the set of emotional terms.

Starting from the assumption that emotional metadata is capable of identifying media items which might be perceived as similar and thus increase the number of relevant recommendations by capturing features across the traditional divide of categories, this chapter splits into two different parts. First it continues the TV path by exploring the possibilities to automatically extract emotional tags (similar to the ones that we saw in *Atmosphere* CS) from the metadata. In the TV domain (or to some extent, video in general) the best metadata for that purpose is the *synopsis* which is almost always present and takes us much closer to the actual meaning of media compared to structured and discrete genres (see Figure 7.1). One of the problems with this approach is that when it comes down to evaluating whether the emotional terms that we extract are valid since we do

not have a user base to validate the results – we can only compare them with synopsis and draw conclusions from that.

The second part addresses this issue by using the Last.fm emotional tags generated by the users. Here we can see if the tags that were extracted automatically from the metadata are the same or similar to the ones that thousands of Last.fm users put manually thus providing a ground truth. Since Last.fm is a social network for music, the second section applies similar methodology as the first one, but on song lyrics.

Both empirical cases presented in this chapter build on the number of publications by the author and published in the period of last 9 months (see Appendixes B, C, D and E). But before we go into these two cases, lets continue the discussion about emotional space.

6.1 Building an emotional space

In the previous chapter we saw that the so-called "unstructured" emotional terms in the TVA *Atmosphere* Classification Scheme seem to have some kind of structure. If this is true then another question is whether the dimensions of such "emotional space" are integral or not. Because if they are integral then, according to Gärdenfors, we can talk about domains eventually leading to concepts.

Even though it seems that whoever designed TVA *Atmosphere* genres was probably not aware of this, people have tried to map out emotions for a number of years. The early works of Rigg and Hevner discussing emotions date more than 70 years back [Rigg, 1937, Hevner, 1935]. Their work has influenced the creation of what we know as *semantic differential* by Osgood, Suci and Tannenbaum 50 years ago [Osgood et al., 1957]. The idea of semantic differential is that not all but most of the emotions can be reduced to three dimensions: valence, arousal and potency. The potency dimension did not stand the test of time and was in many cases excluded from the latter research, leaving us with only two dimensions. If we map out valence and arousal dimension we get something that is called an *affective space* where one axis describes shades of valence from pleasant to unpleasant, and the other defines arousal as emotional states ranging from calm to excited. These two dimensions can be called integral since, much like in a color spindle, we can not assign a value in one dimension without automatically assigning some value on another.

There has been a number of cognitive studies to test and apply semantic differential. One of the bigger ones were conducted by Bradley and Lang in the University of Florida [Bradley and Lang, 1999]. They conducted an experiment where over 200 undergraduate students had to explicitly rate a large number of various english words on these two dimensions – valence and arousal. The results were put together to form the ANEW data set (Affective Norms for English Words).

This seems to be exactly what we need since such emotional space allows us to map out the relative distances between any two affective terms allowing us to make sure that our selected words cover a wide range of emotions and do not end up all being very close to each other.

How should such a space be used? Should we take all the words that it contains? That may be not so effective since first of all there are so many of them and secondly not all of them have the same semantic significanse. Therefore the main idea and the assumption is that we can select certain emotional words in the space and use them as emotional buoys or markers in order to weight the unstructured metadata against those emotional words. By doing that we can transform the synopsis which is unstructured and hard to interpret automatically into the set of terms with numerical values of "how much" each term gets triggered by our metadata. The key is that when we apply this method to a number of TV programs or other kinds of media we always end up with the same set of emotional terms only differentiating by their correlation to the metadata. That means we can much easier compare different programs and see how they relate to each other, drawing on their emotional similarity. See the figure 6.1 for our selection of such affective terms.

We can look at the semantic differential as a way to divide emotional space into four quadrants. In order to cover all of them we selected the terms so that there would be a few of them in each of the quadrants:

- *active/positive* – happy, etc. (arousal: 5-10, valence: 5-10)
- *pasive/positive* – mellow, etc. (arousal: 0-5, valence: 5-10)
- *active/negative* – angry, etc. (arousal: 5-10, valence: 0-5)
- *pasive/negative* – sad, etc. (arousal: 0-5, valence: 0-5)

We wanted to select the emotional terms that were meaningful for people who actually use these words to describe media. We decided to build on the earlier research in this area [Levy and Sandler, 2007, Hu et al., 2007, Collier, 2007], where the authors talked about the formation of the musical ground truth by

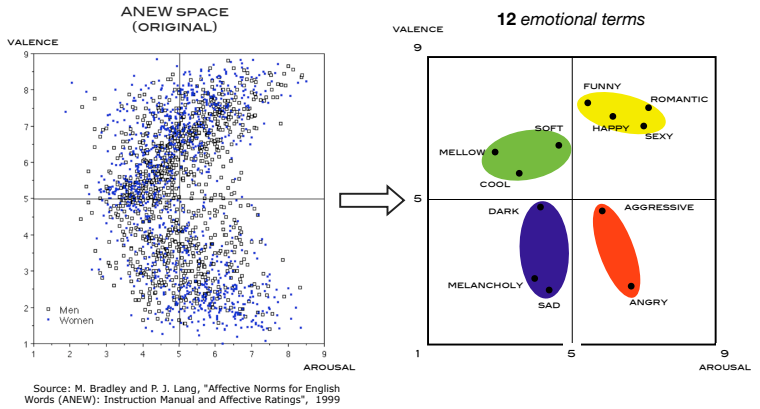


Figure 6.1: The ANEW affective terms mapped out in the emotional space, and the terms that were selected to serve as markers.

presenting the most popular emotional tags that users agreed on in Last.fm social network. Such agreement among thousands of users makes the these words cognitively grounded. We also needed to make sure that selected words are also valid in terms of the selected LSA corpus (in our case, TASA), since we used this machine learning technique for processing the unstructured metadata (synopsis and lyrics).

Choosing the right affective terms is one of the main steps since everything else depends on it. The important thing is to cover a wide enough range of emotions by choosing the words that are representative. Latter in this chapter I demonstrate that expanding our set of 12 terms into the set of 21 does not make a major difference – the emotional patterns are very similar in both cases [Petersen and Butkus, 2008a] (See Appendix D). In the validation section of this chapter I present further discussion on this topic.

Now that we have the terms how do we evaluate our metadata against them? In order to do that we build on the Latent Semantic Analysis (LSA) which is a well known machine learning technique that resembles our cognitive comprehension of text. But first I want to take a step back and to present the two paradigms influencing how we can look into media, and how we end up with the metadata that we have.

6.2 The structure of information

The previous chapter presented an example of how the TV-Anytime annotation works using Classification Schemes. This kind of approach can be thought of as "looking from the outside", because we are not concerned about the inner structure of the media item. Then every media item can be represented as a set of features (for example, genres), leading to relying on feature overlap when estimating similarities between different items. This approach is much simpler to implement since it builds on relatively straightforward statistical techniques, but it is totally dependent of the annotation accuracy since the features are the only things that we can say about the item, and if the features are not precise, then we get a very disturbed image of the meaning of the media.

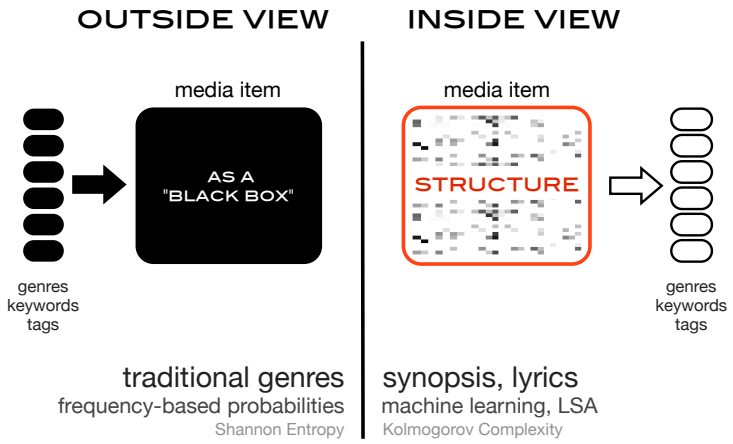


Figure 6.2: Two ways to look at media annotation – an "inside" and "outside" approaches.

Another way to look at the same media item comes from asking *What is the inner structure of the item?*, rather than trying to come up with a set of features first. In this case we may not even talk about the features because we can estimate the similarity by comparing one item's structure to the structure of another item.

These two approaches are backed by a number of scientific theories, ranging from mathematical to psychological. One of the popular ways to look at this is to use the Information theoretical point of view. Information theory was already mentioned in this thesis – in *Chapter 2* I talked about *Information Gain* and

Maximum Entropy methods, while in the *Chapter 5* I used the measure of entropy to express the amount of information. Here I want to turn back to Information Theory as yet another way to approach the problem of finding similar items.

Shannon's approach to information is special because he was only interested in the characteristics of the random information source that transmits objects and not in the objects themselves. In the very beginning of his article "A Mathematical Theory of Communication" Shannon acknowledges that even though messages (or in our context – objects) often have meaning, he states that these "semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages" [Shannon, 1948]. This is one of his main assumptions and clearly defines the scope of Shannon's view at information sources since it simply ignores any internal semantic meaning of the object itself. This is understandable given the historical context of Shannon's theory and the primary goal that he set to achieve. At that time, his theory was focused on improving the transmission of signals over noisy channels, and the optimal coding schemes.

Shannon's view is similar to how we look at media that is annotated using controlled Classification Schemes governed by frequency-based feature overlaps. Here we talk about entropy as a measure of the level of uncertainty of the information source. To compare two items we then can use the measures of *mutual information*, *conditional information* or *information gain* (all these three measure are interrelated), where it all comes down to figuring out how much knowledge about one item reduces the entropy of another item – meaning how much information do the two items share. There is no question that his theory works, the question is only to realize when we should use it due to Shannon's assumptions. If we are dealing with structured metadata, we can employ a number of Information Theoretical measures. This would be a view from "the outside" (see Figure 6.2).

The fact that Shannon ignored the semantic structure of information has been noticed and addressed by R.J. Solomonoff, A.N. Kolmogorov and G. Chaitin in the late 1960s, resulting in the creation "Algorithmic Information Theory" [Solomonoff, 1964, Kolmogorov, 1965] , which uses a notion of Kolmogorov complexity as a measure of information. Kolmogorov complexity focuses on the individual objects themselves where the encoding of an object is a short computer program (compressed version of the object) that generates the original object and then halts. Kolmogorov theory focuses on the inner structure of information object, and this view is similar to what is used in various machine learning techniques where the structure of the item is analyzed to deal with unstructured information. The term "unstructured" here only refers to the fact that the information has not been artificially structured by humans. That means that the

original inner structure is preserved and can be processed latter. That is why I call this a view from "the inside" (see Figure 6.2).

Therefore if our goal is to annotate media item along its quality dimensions, then it is clear that the "outside" approach is often too artificial and in many cases does not lead us to the real quality dimensions, instead it just mirrors various content management side metadata. The "inside" approach to information leads to much more complex processing but it can bring us much closer to the meaning of the content itself.

If we have to choose the machine learning technique that will allow us to take a look into the structure of the media, then we first must think about what kind of information we want to feed into the system. If for example, we want to analyze two images, then we will probably have to pick a method that can analyze the individual pixels, and extract the structure from there. This of course means that we have to be able to analyze raw data to a sufficient level. It is usually not the case in more complex media, like video or audio (although both fields are rapidly developing). If we can not use the raw data itself, then we can settle down for an unstructured metadata that attempts to capture the meaning of media, thus by extracting the inner structure from such metadata, we would get an approximation of the structure of the media itself. Since our metadata is expressed in words, that leads us to choosing Latent Semantic Analysis (LSA) as a preferred method for analyzing such information.

6.3 Latent semantics

LSA [Deerwester et al., 1990, Dumais, 1990] [Landauer and Dumais, 1997] is a well known machine learning technique, used to extract semantic meaning from text. In the information retrieval domain LSA sometimes is called Latent Semantic Indexing (LSI). I want to start by quoting Thomas Landauer (one of the creators of LSA) to show how LSA is different from all the frequency and co-occurence based feature overlaps that we discussed so far: "...the similarity estimates derived by LSA are not simple contiguity frequencies, co-occurrence counts, or correlations in usage, but depend on a powerful mathematical analysis that is capable of correctly inferring much deeper relations (thus the phrase "Latent Semantic"), and as a consequence are often much better predictors of human meaning-based judgments and performance." [Landauer et al., 1998a].

LSA achieves this goal by modeling the usage patterns of words in multiple documents and representing the words and their contexts (documents) as vectors in a high-dimensional space. The frequency at which terms appear and the

context where they occur [1] where they occur are defined in a matrix with rows made up of all the words and columns representing all the documents. Then a single cell contains a number of how many times certain word appears in a certain context. As we can guess, many of the cells contain only zeroes, thus making the matrix extremely sparse. In order to retain only the most essential features the dimensionality of the original sparse matrix needs to be reduced.

One of the ways to reduce dimensionality is by using Singular Value Decomposition (SVD) [Furnas et al., 1983]. SVD decomposes a rectangular matrix into the product of three other matrices (see Figure 6.3). As a result we get the most important dimensions. We can reduce the matrix to a different numbers of dimensions, but empirically the most efficient one seems to be around 100 – 300 dimensions (see Figure 6.4). SVD makes it possible to model the semantic relatedness of paragraphs and terms as vectors, with values towards 1 signifying degrees of similarity between the items and low or minus values typically around 0.02 signifying a random lack of correlation. In this semantic space paragraphs or words which express the same meaning will be represented as vectors that are closely aligned, even if the words do not literally appear in any of the same documents. Instead these terms may co-occur in other documents describing the same topic, and when reducing the dimensionality of the original matrix we can discover the latent relationships between words or documents.

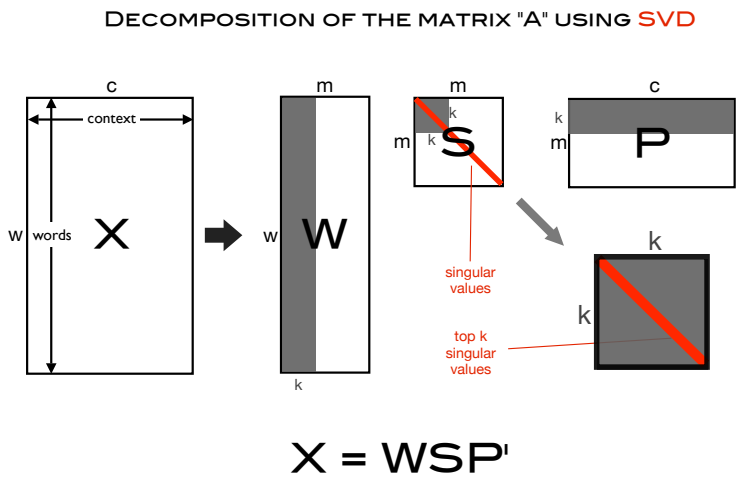


Figure 6.3: Linear decomposition of the matrix X using SVD.

[1] in LSA a context is simply a piece of text where the word happens to occur. Such piece of text can refer to a single essay, but it may also be reduced to a single paragraph, the sentence, or even be expressed as a sliding window (lets say 15 words before and after the word). Naturally the smaller our window is, the more sparse the *term-document* becomes

A collection of documents in LSA form a text corpus. There are a number of text corpus in the world based on different languages or on different domains of knowledge. We chose to use the standard TASA text corpus from Touchstone Applied Science Associates, Inc., consisting of the 92,409 words found in 37,651 texts, novels, news articles and other general reading material that American students are exposed to up to 1st year college.

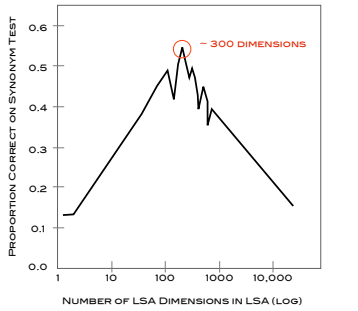


Figure 6.4: In a number of empirical test it was determined that LSA produces best results when the SVD takes only the first 300 dimensions.

The schematic view of what happens when we feed the text (in this example, TV synopsis) into LSA and try to mach it against the one of the emotional terms is outlined in the figure 6.5. This example also shows our empirical setup provided by the University of Colorado at Boulder (http://lsa.colorado.edu) for calculating cosine similarities in order to project TV synopsis into emotional space.

LSA can be used in a number of different ways to calculate various things: finding the nearest neighbors for a given word, comparing one piece of text to another, etc. We used LSA to do a "one to many" comparison task. This means we had two kinds of input: our 12 affective terms (in a figure 6.5 denoted by A) and a piece of text representing the synopsis (denoted by B). The only parameters that can be chosen in a given setup were the selection of the actual text corpus (in our case TASA), and the selection of dimensionality from a range of 0 to 300 (we sticked to the default 300). What we want to get in the end is a correlation value between the synopsis and every of the affective terms.

What happens is that both synopsis and the affective term are represented as vectors in a word space showing the direct occurrences of the words in the corpus. After that LSA uses a preprocessing step called *weighting*. "LSA,

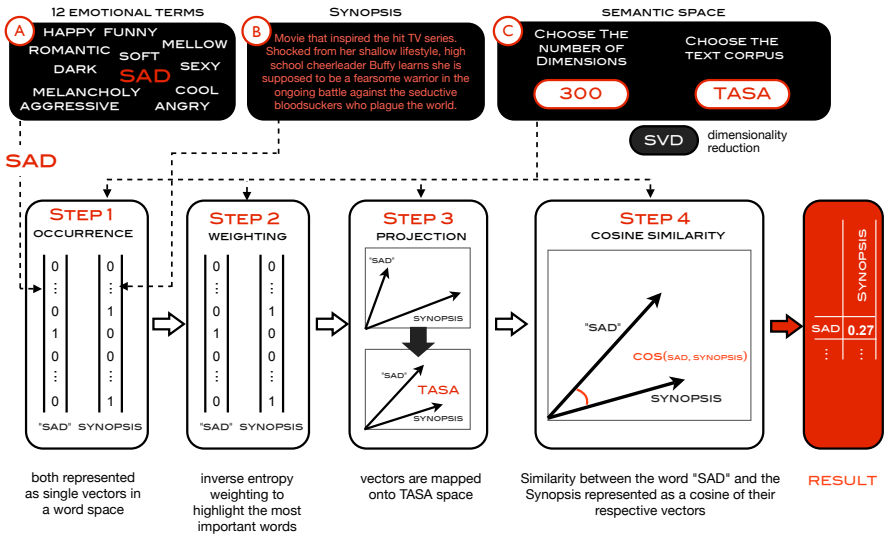


Figure 6.5: A process of calculating cosine similarity between synopsis metadata an emotional term using LSA.

unlike many other methods, employs a preprocessing step in which the overall distribution of a word over its usage contexts, independent of its correlations with other words, is first taken into account; pragmatically, this step improves LSA's results considerably" [Landauer et al., 1998a]. This is done by using the inverse entropy weighting to highlight the most important words. After that the weighted vectors are mapped onto TASA space where a position of the vector is based on a latent knowledge of how the words relate. Once we have both vectors projected onto a space, then we can calculate cosine similarity (see Chapter 2.4). Such procedure is then repeated for every one of the 12 affective terms, which eventually leads u to having the projection of the synopsis into emotional space based on LSA.

6.4 Building concepts using emotional terms

In the previous chapter I highlighted that the main obstacle for building valid conceptual spaces is the fact that we do not know the quality dimensions of the TVA Classification Schemes, confirming exactly what Gärdenfors identifies to be the main challenge. This made it complex for the Classification Schemes to be called domains the way they are understood in the Conceptual Spaces theory.

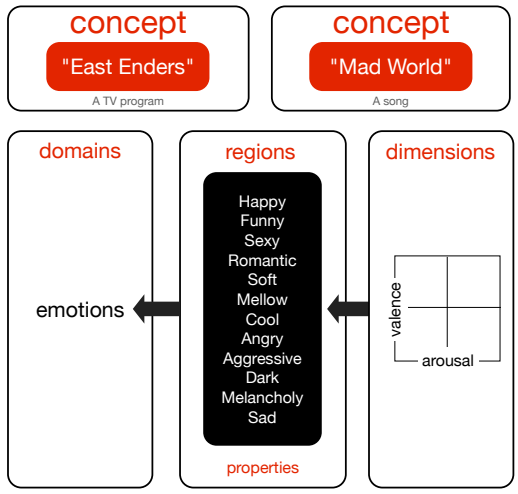


Figure 6.6: An example of a TV program and a song represented as a concept by combing dimensions, properties and domains

Nevertheless we saw the potential of the *Atmosphere* Classification Scheme since it could be represented as a space. We could not do that with the raw genres due to their structured form and the fact that they are meant to be used as they are, and are already stripped from most the latent semantic information.

Such latent information is very much alive in the unstructured metadata, such as a TV synopsis or song lyrics. I still stick to the emotional point of view because it is precisely how most people interact with media, and emotional similarity is known to cross traditional genre boundaries leading to more novel recommendations.

Therefore I propose that media can be represented as concepts using the cognitively grounded emotional terms extracted from unstructured metadata (see Figure 6.6).

In this case the dimensions can be defined using the notion of *semantic differential* where every basic emotion can be mapped in a two dimensional space according to their valence and arousal values. This is very much grounded by a number of cognitive experiments and what is even more important these two dimensions are integral. This allows us to talk about emotional space as a domain where similar emotions can be grouped and distances between them can be calculated (or at least estimated).

When I was talking about TV program concepts based on Classification Schemes, then there were 5 different spaces resembling the "domains" - even if these were not well formed integral domains. One of them was the *Atmosphere* domain which actually looked like the only valid domain. In this chapter I have only one domain to begin with since I chose to look at media only through the emotional lens, basing this on the assumption that emotional value is one of the main components of media. The importance of emotions for media is quite different in TV and music cases. In TV we have many more angles we can take when we look at media (for example, what is the content, intention or format of the program, etc), while in music it is really very much down to the emotions. This means that the presented methodology does not aim to give a complete picture of the meaning of a TV program or a song, but instead it extracts the fraction of a meaning from the media based on emotions. This fraction is much larger when we talk about songs, constituting to the majority of the meaning, while in a TV domain such fraction is considerably smaller. But even there it is very valuable since it brings an emotional angle into the media recommendation process.

This means that when we talk about the concept of a certain media item – TV program or a song – we must consider all the domains in relation to each other. In this chapter I am taking only emotional domain and ignoring all the other ones, thus even though the conceptual picture is not complete, it shows one of the main components of media – emotional value.

Based on the *semantic differential* we can divide the emotional domain into four main regions – *Active Positive*, *Passive Positive*, *Passive Negative* and *Active Negative* – where all of our smaller regions (individual affective terms) can be with a certain precision mapped into one of the four bigger regions. This serves as different level of abstraction. Sometimes when we can not pinpoint the exact emotions, it is still valuable to be able to determine which one of the four bigger regions come out as more active pointing us to the general direction of what emotional balance we can expect from this song or a TV program.

The following two sections (6.5 and 6.6) present the data from two cases: the *TV* and the *Music* where the affective terms are extracted from respective metadata fields to represent the emotional value of a given media.

6.5 TV program personalization using emotional terms

The first case deals with TV programs since it builds on the previous chapter which allows us to see how different and descriptive our emotional information is compared to the traditional genre-based annotation. So far we have our semantic space of emotions, 12 affective terms in that space, synopsis descriptions for each programs as representatives of unstructured metadata that will allow us to approximate the meaning of the content itself, and finally we have LSA as method for extracting latent similarities. How it all fits together can be seen from the Figure 6.7. LSA here is presented as a "black box", since its structure was already presented in a section 6.3 (see Figures 6.3 and 6.5).

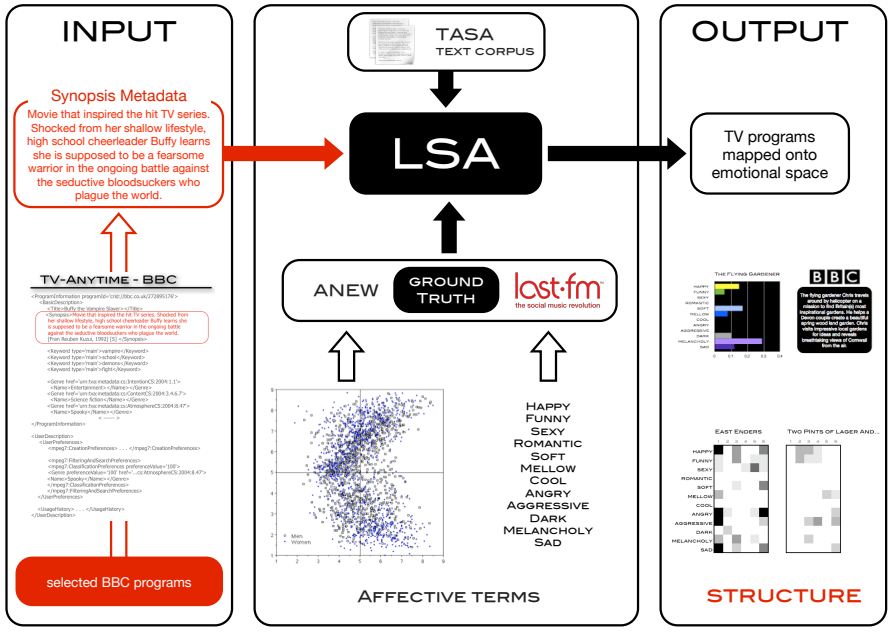


Figure 6.7: An overview of the methodology to extract emotional terms from the TV synopsis

Lets see how this methodology can be applied on TV synopsis data and what results come out of this. First we selected eight BBC programs, most of them are the ones already presented in the previous chapter. Then we took the synopsis information from each of the programs and computed the cosine similarities

	Buffy	Super Vets	The Flying Gardener	Newsnight	I'm Boy Anorexic	Ready Steady Cook	Two Pints of Lager	East Enders
HAPPY		0.05	0.07					0.18
FUNNY				0.11				0.1
SEXY	0.06			0.2				0.09
ROMANTIC		0.12				0.15		
SOFT			0.13					0.17
MELLOW	0.06		0.1				0.09	
COOL			0.17					0.06
ANGRY							0.12	0.34
AGGRESSIVE	0.09	0.06	0.12				0.14	0.11
DARK	0.08				0.07			
MELANCHOLY			0.05				0.05	
SAD		0.07		0.06			0.12	0.17

Figure 6.8: LSA cosine similarity between the synopsis descriptions of programs "Buffy the Vampire Slayer", "Super Vets", "The Flying Gardener", "Newsnight", "I'm a Boy Anorexic", "Ready Steady Cook", "Two Pints of Lager And" and "East Enders" based on their 12 frequently used last.fm affective terms.

between a synopsis text vector and each of the selected last.fm emotional words (see Figure 6.8). One final note here before we get into the actual data. At first it may sound strange that the affective terms are taken from the last.fm even though this is a TV domain. This was done assuming that to a certain level emotional response expressed by the users in last.fm in relation to songs, would represent their response to the emotional context of a TV program. A huge user base consisting of hundred thousands of Last.fm users gives a good indication of a general ground truth and shows what are the emotional words that people agree on. Since to the authors knowledge no such amount of free emotional tag data exists for a TV domain, the music domain was used in return. The following results indicate that the approach is accurate enough to infer the general emotional feeling of a TV program. The table shows only the correlation values that do not go lower than 0.05 since the lower correlation may be perceived as noise.

What can be said when we look at the results table? What we are looking for, is to see which emotional terms appear to be more dominant than others thus setting the emotional tone for the whole program. To see the dominant terms more clearly the information from the table was translated into graphs. In the following pages you will find a presentation of graphs representing the emotional program balance coupled with their respective synopsis metadata fields, so that we could see if the affective terms capture the overall emotion of the program or not.

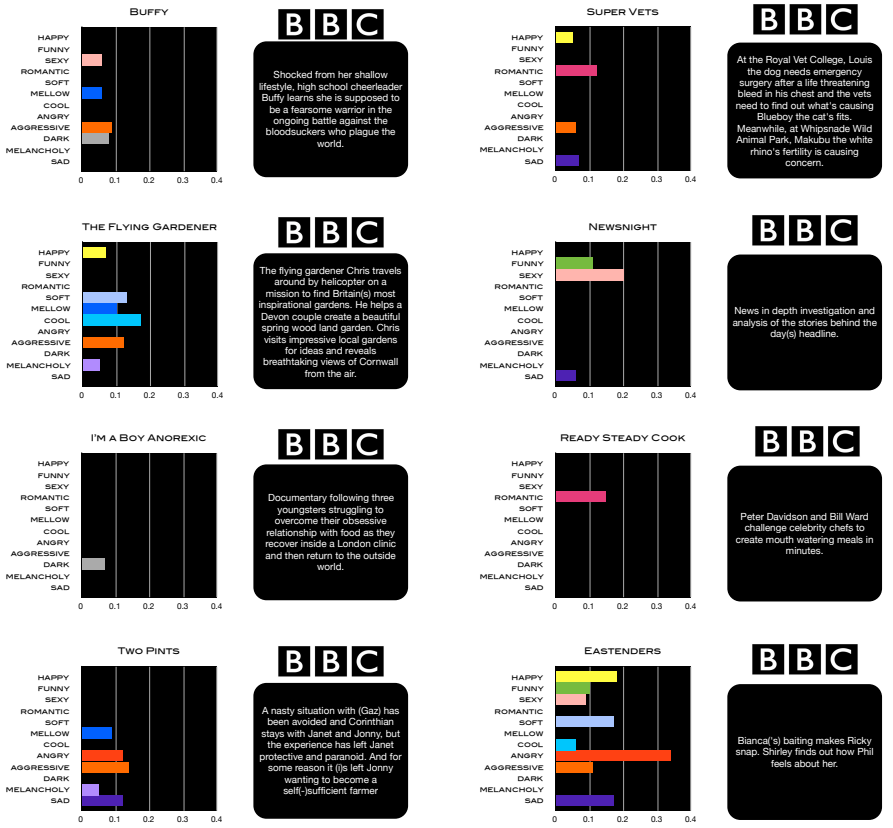


Figure 6.9: LSA cosine similarity between the synopsis descriptions of 8 BBC programs and the 12 frequently used last.fm affective terms.

In the LSA analysis of the program "Super Vets" (see Figure 6.9) we see a strong emotional contrast: *Happy* vs. *Sad*, *Romantic* vs. *Aggressive*. This can be explained by looking at the synopsis where the unpleasant part comes from the the animal being sick, having operation, bleeding, and in general appearing in a life threatening situation. Whereas emotional intensity is stressed by the term *Romantic*. The presence of *Happy* shows that the correlation between the synopsis and the chosen tags might often trigger both combinations of complementary elements as well as contrasting emotional components rather than a single monochrome feeling. Similar situation is also in the next program "The Flying Gardener" where we find a broader emotional spectrum. The synopsis triggers a concentration of passive pleasant valence elements related to the words *Soft*, *Mellow* combined with *Happy*. In this context also the tag *Cool* comes out

as it has a strong association to the word "air" contained in the synopsis, while the activation of the tag *Aggressive* appears to be less explainable.

In a program "Buffy the Vampire Slayer" we see four dominant emotions extracted using the affective terms: *Sexy*, *Mellow*, *Agressive* and *Dark* with the two later ones being more dominant. We can see that it fits the synopsis description quite accurately. Emotions like *Dark* and *Aggressive* are triggered by the synopsis talking about fearsome warrior, ongoing battle and bloodsuckers who plague the world. Emotion *Sexy* most likely comes from Buffy being a cheerleader. The two terms *Sexy* and *Dark* fit the actual TV show perfectly thus capturing the general tone of the program.

An analysis of the program "News night", based on the short description triggers the tags *Funny* and *Sexy* which might not immediately seem a fitting description, probably caused by these emotional terms being directly correlated with the occurrence of the words stories and news within the synopsis.

Only a singular emotion can be retrieved from the documentary "I am boy anorexic" which is *Dark*. This is a very understandable since the synopsis is filled with the words like "struggling", obsessive relations", "recover" and "clinic". Another singular emotion program is the lifestyle program "Ready Steady Cook!" which triggers the tag *Romantic* as associated with meals.

Next we turn to what should be a comedy – "Two Pints of Lager And..." [2] – but it comes out as filled with lots of negative emotions, which is an expected outcome given the synopsis. This raises a question how reflective a synopsis for a single episode is when we are trying to estimate the emotional balance of the whole program. Final program is a soap "Eastenders". It is interesting to see so many different terms being triggered by such a short synopsis. This looks very much what you would normally expect from a soap – lots of contrasting emotions resulting in a drama.

All these synopsis so far were based on a single episode. Would the picture change once we sum up the emotional term values from a number of episodes over a period of time? To answer that we need to find the programs that had different synopsis for every single episode. We found out that there are not that many of those. Eventually we ended up with one illustrative example of two programs with changing synopsis for each episode – "Eastenders" (a soap) and "Two Pints" (a comedy).

[2] I will use the short version of the title – "Two Pints" – to refer to this program.

First we need to get more synopsis descriptions. We took 6 consequent episodes of each program giving us the following synopsis (see Figures 6.10 and 6.11)

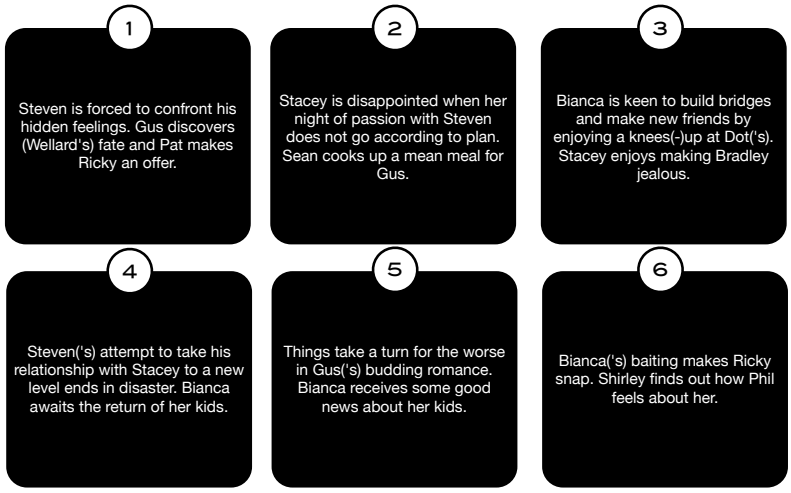


Figure 6.10: 6 synopsis of a soap "East Enders"

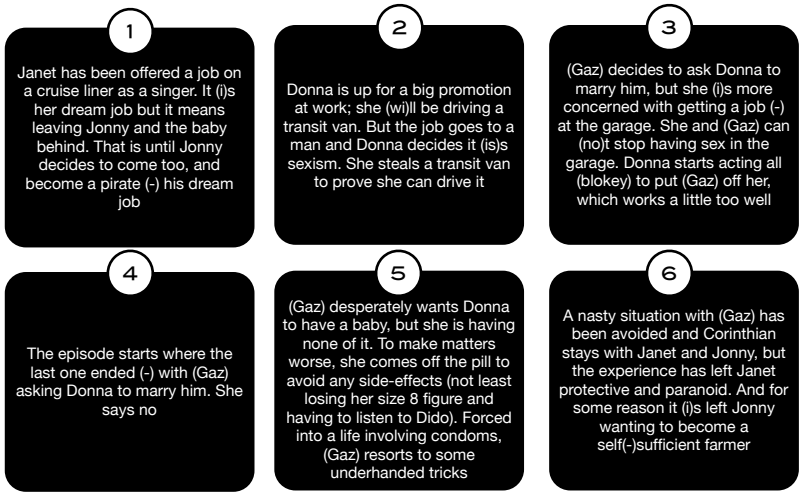


Figure 6.11: 6 synopsis of the comedy "Two Pints"

Once processed these synopsis make up the following tables where we see which affective term are triggered by each of the episodes (see Figure 6.12). Here you can see all the values, even the ones that go bellow 0.05 which we consider as noise. These ones will be removed when we sum up the numbers to make the graphs representing the total emotional balance of a show (see Figure 6.13). First of all we see that "Eastenders" emotional distribution remains more or less the same as in the figure 6.9, whereas "Two Pints" changes quite a lot. In fact it changes to what we would expect from such a program since the earlier dominant values of *Angry*, *Aggressive* and *Sad* are now replaced with *Happy* and *Funny* – very natural emotions for a comedy to have. *Aggressive* emotion seems to come out here as well, which much like in "The Flying Gardener" is somewhat strange and unexpected.

EASTENDERS

	1	2	3	4	5	6
HAPPY	0.3	0.07	0.16	0.03	0.02	0.2
FUNNY	0.1	-0	0.16	-0	0.09	0.1
SEXY	0.1	0.01	0.03	0.09	0.2	0.1
ROMANTIC	0.1	-0	0	0.01	0.04	-0
SOFT	0.1	0.02	0.1	-0	0.04	0.2
MELLOW	0.1	0.07	0.05	0	0.07	0
COOL	0	0.04	0.01	0.01	0.02	0.1
ANGRY	0.6	0.1	0.04	0.05	0.09	0.3
AGRESSIVE	0.3	0.06	0.05	0.02	-0.1	0.1
DARK	0	0.12	-0	-0	0.02	-0
MELANCHOLY	0.1	0.05	0.15	0.07	0.09	-0
SAD	0.5	0.02	0.11	0.08	0.08	0.2

TWO PINTS

1	2	3	4	5	6
0.1	0.14	0.13	0.1	0	-0
-0	0.13	0.1	0.1	0.03	-0.1
-0.1	0.08	0.04	-0	0.04	0
-0	0.01	0.08	-0	0.01	0
0.1	0.02	0.01	-0	0.08	-0.1
0	0.02	0.01	-0	0.12	0.1
-0	-0	0.01	-0	0.02	0
0	-0	0.03	0.03	0.02	0.1
-0	0.07	0.11	0.16	-0.1	0.1
0	0	-0.1	-0	0.01	-0
0	-0.1	0.11	0.03	0.06	0.1
0.1	0.04	-0	-0.1	-0.1	0.1

Figure 6.12: LSA cosine similarity of the soap "East Enders" and the comedy "Two Pints" synopsis with the 12 affective terms over six episodes.

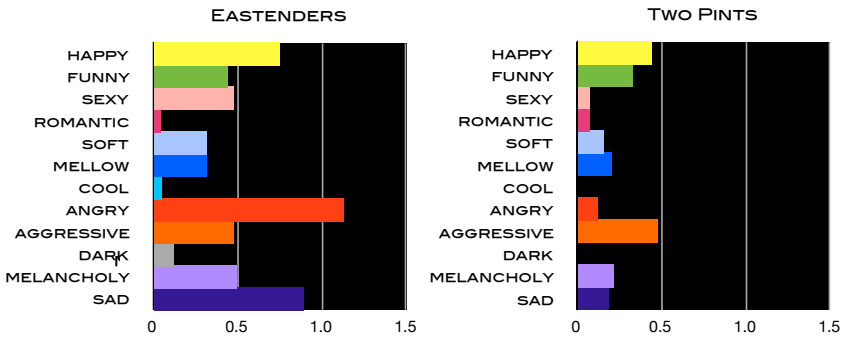


Figure 6.13: The accumulated correlation values for both programs.

These were the accumulated graphs. We see that it is useful to take more episodes into consideration since it tends to even out the peaks and lows of an individual episode and gives a much clearer picture of what the program itself it all about. How about if we take the same six episodes but instead of summing them up, we will simply map them out one episode after another. Here is where we get the pattern of emotions. To visually indicate the correlation between the synopsis and each of the emotional terms a color scale has been chosen where white color indicates no correlation and black color shows the strongest correlation (since the correlation is based on the cosine between two vectors then 0 shows no correlation while 1 shows the exact match). The scale is linear and ranges from 0.05 to 0.23 (10 different shades of gray) (see Figure 6.14).

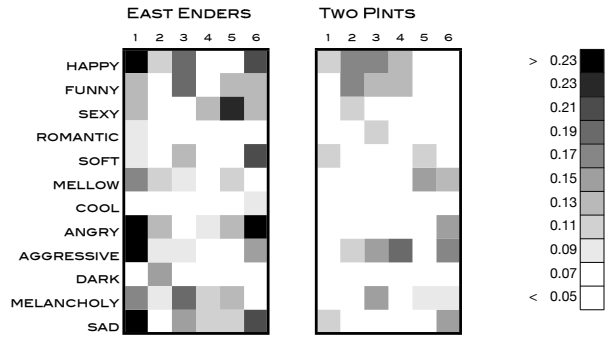


Figure 6.14: LSA cosine similarity of the soap "East Enders" and the comedy "Two Pints" against 12 frequently used last.fm affective terms accumulated over six episodes.

Without going into too much detail, we see several differences between these two patterns. In the soap "Eastenders" synopsis emotional terms are triggered much more often and they are also much more saturated, whereas in the comedy "Two Pints" seems to much lighter and much more sparse. Even though we have not gathered enough empirical data to make a strong statement here, we propose that even program types could be differentiated by projecting the synopsis into emotional space and comparing the patterns (see Appendix B). This is where the TV case stops. I will get back to it again latter in the end of this chapter for the validation and in the discussion chapter. So far we see that even the very short piece of text contains latent emotional information that can be extracted using the set of emotional markers and applying LSA as machine learning method. This can be credited to the fact that synopsis information while being unstructured is meant to describe the essence of media, whereas the annotation using genres only attempts to categorize programs into a set of predefined categories.

How much closer can we get to the content without going into audio and video signal processing? In the TV domain synopsis is probably the best candidate, unless we have a script of the program, which is not usually the case. The synopsis metadata gets us as close as possible while still remaining only metadata and not being an integral part of the content itself (see Figure 7.1). Since this thesis focuses on the two types of media – audio and video, I want to present the final case, where a very similar methodology is applied to music. Even though it is not very common, but songs may also have short textual descriptions corresponding to what we could call a synopsis. If this is not available, then the Internet is full of various reviews, that can also be considered a synopsis-type of metadata. But instead of going the same way as I did with the TV programs, I would make a shift and try to get even closer to the actual meaning of the media. It was pointed out in the *Chapter 3* that the song lyrics, for example, can be considered as a metadata and a content at the same time. My assumption is that if we can process lyrics in a similar fashion as we did with synopsis metadata, then it would represent a structure much closer to the actual structure of the whole media item (a song) rather than just the structure of the metadata.

6.6 Music personalization using emotional terms

Emotions in music are dynamically unfolding in time. Over the past half century these aspects of musical affect have been the focus of a wide field of research ranging from how emotions arise based on the underlying harmonic and rhythmical structures forming our expectations [Meyer, 1957][Huron, 2006], to how we consciously experience these patterns empathetically as contours of tensions and release [Jackendoff and Lerdahl, 2006]. This in return triggers physiological changes in heart rate or blood pressure as has been documented in numerous cognitive studies of music and emotions [Krumhansl, 2002].

Recent studies suggest that musical structure to a much larger extent than previously thought is being processed in "language" areas of the brain related to temporal structure and construction of meaning in general evolving over time [Levitin and Menon, 2003]. Specifically related to songs both fMRI and ERP neuroimaging experiments point to linguistic and musical dimensions as being processed by similar overlapping brain areas. This seems to support the hypothesis that the linguistic and melodic components of songs are processed in interaction [Schön et al., 2005] and are not isolated from each other. Experiments indicate that song memory is not organized in strict temporal order, but rather that text and melody intertwine based on the connections of higher order structures [Peretz et al., 2004].

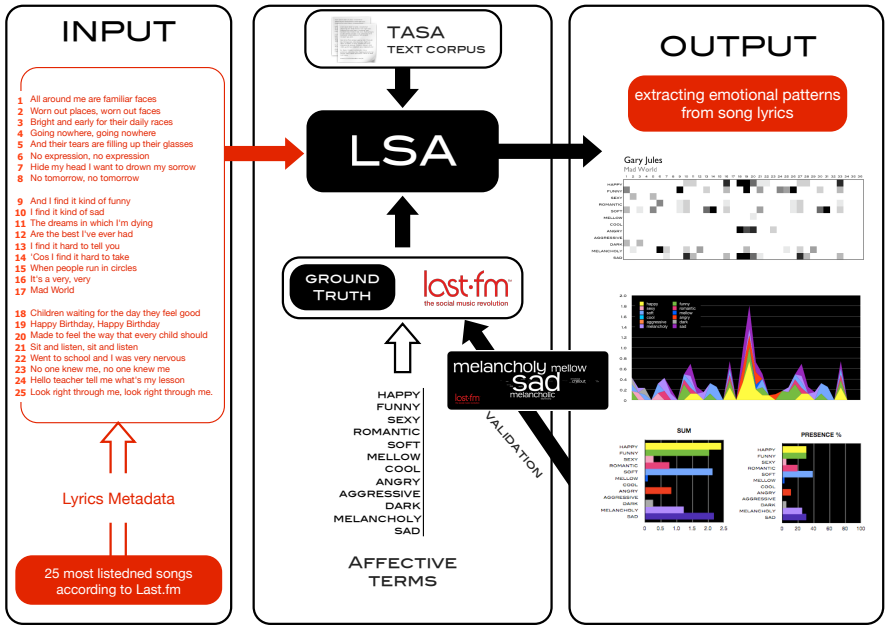


Figure 6.15: An overview of the methodology to extract emotional terms from the song lyrics

The empirical setup for the song analysis is very similar to the one used to analyze the TV programs. The main difference is that in the TV case we took the synopsis as a whole resulting it being represented as a single vector in the semantic space, whereas when processing lyrics we took every line individually resulting in having as many vectors as there are lines in the song. By calculating the cosine correlations between those vectors and the 12 vectors representing each of the affective terms we end up having a 2D pattern for a song where vertically we have 12 affective terms and horizontally the lyrics of the song plotted one line at a time. Since the lyrics lines appear one after another and are highly related to each other we can observe patterns or shifting contours of emotions throughout the song (see Figure 6.15).

Following the process presented in the figure 6.15 we end up having one pattern for each song. There are several ways how to interpret them. First, we can look at the pattern itself by looking for the structure expressed as individual pixels grouping together. The vertical axis in the pattern has all of the 12 affective terms grouped into more basic catgories: at the top we have the *active positive* emotions represented by the terms *Happy*, *Funny*, *Sexy* and *Romantic*, after these we then have the *passive positive* terms Soft, *Mellow* and *Cool*, then we

move into the negative emotions starting with the *active negative* (*Angry* and *Agressive*) and finishing with the *passive negative* (*Dark*, *Melancholy* and *Sad*). The horizontal axis is *time* (in reality this axis represents the number of lines in the lyrics, but it can serve as an approximation of time) which in our case ranges from as few as 19 lines to as many as 50. The strength of the correlation between the single affective term and single lyrics line is represented by different shades of gray – ranging from white (no correlation) to black (strong correlation). In order to keep the noise level down we chose to discard values of correlation between lyrics and tags if they got below a threshold of 0.09 [3].

In this thesis all song analyses follow the same template. You will notice the *12 affective terms / song lyrics* pattern at the very top followed by three graphs under it showing different aspects of emotional distributions over time, and two bar-graphs showing the summarized values of the affective terms. At the top right corner you see a table containing four columns: 1) the number of times the affective term correlation with the lyrics line going over the necessary threshold; 2) the sum of all values for each of the affective terms; 3) the average of all the "above the threshold" correlation values for each affective term; 4) the number from the first column divided by the number of total lines that the song has, expressed as a percentage. The row at the very bottom of the table shows the average values of the respective columns (the average value of the fourth column shows the overall emotional presence in the the song). The rest of the information is expressed in 5 different graphs. Before going to the first example (*Metallica – Nothing Else Matters*), first I will explain what each of the graphs show.

Instead of looking at the raw pattern we can project a song into one of the axis – the affective terms (vertical) or the time (horizontal). Each projection presents different aspects of the song and in the overall evaluation have to taken in combination (Figure 6.16).

The vertical axis allows us to see the summarized values of every affective term for a song. We can do the summarization in a number of different ways. First we decided to do the same as we did in the previously presented TV case where we simply added al the values (above the threshold) and ended up having one value for every one of the 12 terms. Once plotted in a bar-graph these 12 values show us which affective terms come out the strongest over the course of the song. We have noticed certain distributions that such graphs tend to form (see Figure 6.19).

[3] We found that using the lower threshold values, for example 0.05 like in the TV case, generates too much noise.

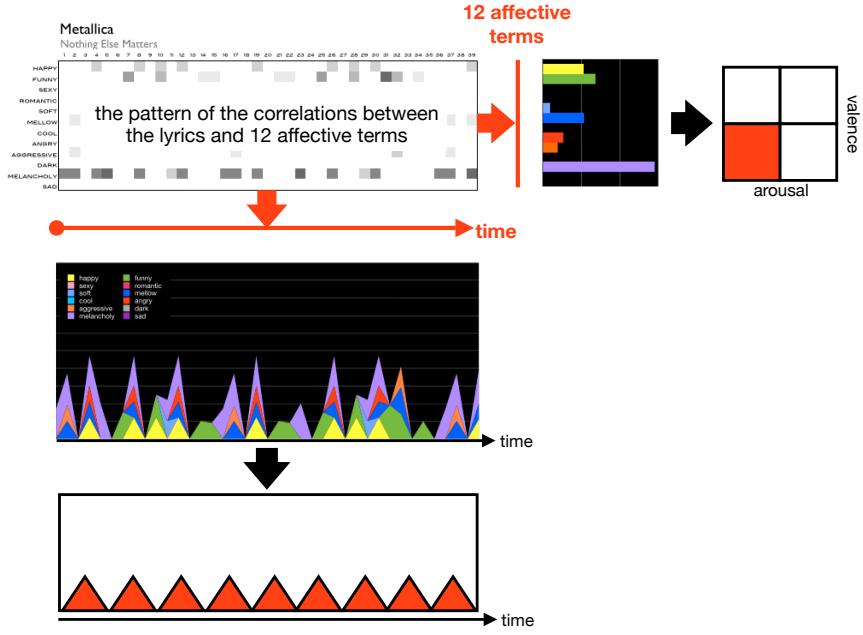


Figure 6.16: The pattern that we get by calculating the correlations between 12 affective terms and the individual lines of the lyrics can by projected to the horizontal or vertical axis, highlighting either accumulated values of the affective terms of emotional distribution over time.

What you can not see from the sum of the individual values is how often a certain affective term appears in the song. For example, in the song "The Pretender" by Foo Fighters (see Figure F.19) the term *Dark* gets triggered only by five of the song lyrics lines from the total number of 50. But every time it comes out, it does so by having on average a very strong correlation to the lyrics line (correlation average = 0.45) thus in the end resulting by accumulating to a very high value despite appearing so few times (accumulated correlation = 2.3 – 2nd strongest from the 12 terms). Affective term *Funny* in the same song comes out 18 times (out of 50 total lines), but since it triggers on average much weaker correlation (correlation average = 0.13) it accumulates to the total value of 2.4, being just slightly higher that the term *Angry*. Does that mean that this song is just as *Funny* as it is *Angry*? To answer that we can first of all look at the neighboring values of *funny* and *dark* to see which one gets more support, or we can take into consideration that one appeared 18 times (36% of the whole song) and another only 5 times (10%). To see the distribution of the affective terms based on the presence of each term over the course of the song we calculated how many times each term get triggered and plotted it in another bar-graph

showing how often each term appears (expressed as a percentage of the total length of the song). When taken together these two graphs give a clearer picture of the distribution of the affective terms for a given song. In most cases both *SUM* and *PRESENCE* graphs look similar thus reinforcing each other, but in a few cases (for example Foo Fighters – "The Pretender") graphs allow us to see which measure captures the emotional value of the song the best (compared with the last.fm tags) – accumulated values of the affective terms or simply their presence.

The horizontal axis in our emotional pattern is *time*. It allows us to monitor the changing contours of emotions over time. In the TV case the time dimension was too weak to be taken into consideration [4], whereas in music it carries a lot of information. I have not performed in depth analysis on the patterns based on the time axis yet, but several things can be already noticed.

The first graph right under the original pattern show the accumulated sum values of each affective term over the course of the whole song. On the left we start from 0 and on the right we get the exact same values as in the first bar-graph (*SUM*). The value of this particular graph is that we can see how the affective term correlations add up over time – we can evaluate how steep the certain portions of the curves are showing us the strength of the single correlation value. Also we can judge the overall presence of emotion by looking if the curve goes up most of the time instead of just jumping once in a while. In principal this graph is does show anything new compared with the original pattern, but it presents information in a different way making it easier to see certain features of the pattern. The scale for this graph is different for each song, since different songs have different maximum correlation values.

The second graph also shows all the correlation values, but instead of adding them up it simply presents them separately on the scale from 0 to 1 (the same for all the songs). This graph makes it easy to see in what range the single correlation values are and help us to decide what kind of scale to chose to the main pattern [5]. It also shows very clearly what is the maximum correlation value that any of the affective terms take in this song. At the upper right corner of this graph you can see the maximum value of all the correlations (*MAX*) and the average value of all the correlation values taken together (AVERAGE) [6].

[4] Even though previous TV episode may (and often do) have direct influence on the next one, we perceive them separately because of the time difference between them being days or even weeks whereas in music the time difference between lines is a few seconds

[5] The scale was chosen to have 10 grades of 0.02 each ranging from 0.09 to 0.25.

[6] The average value shown in the 2nd graph (*AVERAGE*) is different from the (*avg*) shown in the table in the upper right corner of the template. This is because the first one is the average correlation of all the values in the table (the ones that are 0.09 or higher) whereas the second one shows the average of the 12 average values of each affective term individually

The third graph shows the accumulated values not for each of affective terms individually but for all the 12 terms together. Instead of looking at the separate affective terms individually, here we see the sum of all the correlations between a single lyrics line and all 12 affective terms. This graph helps to notice the lines in the lyrics that are very "emotionally charged". We can observe certain patterns characterized by the number of peaks and their place in time. One could call it some kind of "emotional energy" bursts. The scale for this graph, just like the first one, is also different for each of the songs.

Finally at the lower right corner of the template you see the black square with the last.fm logo presenting the emotional tagcloud for a given song retrieved from the last.fm social network. First the complete tagclouds for each song were retrieved using last.fm API (Application Programming Interface) [7]. Since only a fraction of the tags have emotional value (it differs from song to song), the emotional tags were manually selected and used to generate an emotional tagcloud. This was done in order not to overshadow the emotional tags with some of the more popular yet much less meaningful tags as *Rock*, *Seen Live*, *80s*, etc. The last.fm emotional tagcloud is used for validating our results and evaluating how similar our automatically extracted emotions from song lyrics are compared to what thousands of last.fm users say about the songs.

As the empirical data 25 songs were selected from the list of the most popular songs given by last.fm social network (see Figure 6.17). The reason behind selecting these particular songs is that they are popular enough to have accumulated significant number of user generated tags allowing us to compare our results to the opinions of last.fm listeners. This is indeed a very small dat set therefore I will be very careful by drawing any strong conclusions from our results. What we can see are some general tendencies and we do need more data to validate our method. The main reason why we are using so few songs so far is the fact that we used the LSA setup provided by *The University Of Colorado* which is implemented as a web interface therefore limiting us from automating the process. Recently we have finished building our own LSA setup which gives us many more options since we now have the full control over the algorithms and the text corpus (for some early results about our own text corpus see *Chapter 7* and the *Appendix E*). Several songs are presented in *Chapter 6* and *Chapter 7* while the rest of the song are given in the *Appendix F*.

[7] http://www.last.fm/api

ALANIS MORISSETTE	21 things i want in a lover
THE KOOKS	always where i need to be
LEONA LEWIS	bleeding love
COLDPLAY	clocks
COLDPLAY	the scientist
COUNTING CROWS	colorblind
NORAH JONES	come away with me
RADIOHEAD	creep
R.E.M	everybody hurts
GLEN HANSARD	falling slowly
GOO GOO DOLLS	iris
GARY JULES	mad world
EVANESCENCE	my immortal
METALLICA	nothing else matters
FEIST	now at last
AMY WINEHOUSE	rehab
JOHNNY CASH	san quentin
NIRVANA	smells like teen spirit
LED ZEPPELIN	stairway to heaven
MUSE	starlight
THE POSTAL SERVICE	such great heights
FOO FIGHTERS	the pretender
MGMT	time to pretend
LINKIN PARK	what i've done
OASIS	wonderwall

Figure 6.17: A list of 25 songs used in this thesis to illustrate the potential of the proposed methodology to extract emotional terms form song lyrics.

Let me present an example of a song analysis illustrating the use of different measures and what can we tell from them. The song for this example is "Nothing Else Matters" by Metallica. After calculating the correlation between the 12 affective terms and the song lyrics we end up with a pattern which is presented at the top of the template. From it we notice the lack of any strong correlations between the affective terms and the lyrics – the pattern looks very light. From the table on the right to the pattern we can see that only 13% of the effective area is triggered by emotional words (in relation to other 24 songs this is an average result with numbers varying from as low as 8% to as high as 34%). The only row that stands out in the whole pattern is *Melancholy*, which is both saturated – indicating the strength of the individual correlations (average value = 0.17); and steady – indicating the constant presence of *Melancholy* emotion (appears in the 44% of the song).

The *SUM* bar-graph on the right shows the total dominance of the single emotional term *Melancholy* and if we need to abstract to a higher level it would result in having only one dominant region – *passive negative*. The second bar-graph reinforces the first one by highlighting the dominance of *Melancholy*.

The first graph under the original pattern shows the *Melancholy* steadily raising from the very beginning of the song to the very end, while all the other emotions

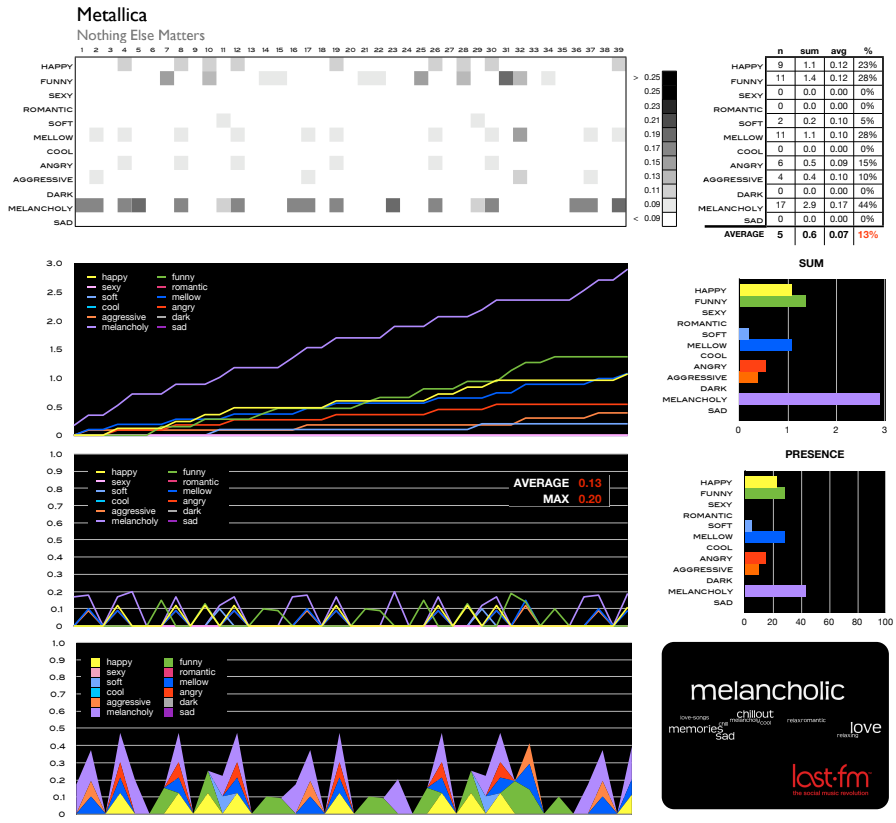


Figure 6.18: The emotional pattern analysis of the "Nothing Else Matters" song.

appear to be dimmed. The second graph shows all correlation values plotted individually. We can see that there are no high peaks and that in most cases *Melancholy* gets the highest correlation value compared with other terms. The third graph at the bottom simply confirms all the other graphs by showing very steady accumulated emotional pattern in relation to time. Partly this is because of the very dominant single emotion – it is so strong in relation to the other ones that it dictates the shape of the accumulated pattern as well. Since *Melancholy* in this particular song is very steady appearing almost every second line the that means that it overshadows all the other emotions constantly from the beginning till the end. And finally because the *Melancholy* has very little fluctuations in terms of its correlation value (once it appears it always seems to be at the similar level) the pattern does not change much over time.

Looking at all the six graphs and the table we can conclude that without any doubt the *Melancholy* is the most dominant emotion reflected in the song lyrics. It matches perfectly with what the last.fm tagcloud suggests – vast majority of the listeners tagged the song as *melancholic*. In order to perform more accurate validation we have to abstract both results (our pattern and the last.fm tagcloud) to a higher level categories since the actual words used are different. The validation of this song is presented in the following section (see Figure 6.32).

I already mentioned that once we project the *lyrics / 12 terms correlation* pattern into a vertical axis by accumulating the rows we end up having a bar-graph where certain distributions can be noticed. Let me elaborative a little bit on that presenting some of the early results that we got from analyzing 25 songs.

One way to validate our results is to abstract our bar-graphs into more basic emotional categories represented as separate regions in the emotional space. Then we can do the same abstraction on the last.fm emotional tags and see if the two match. Of course in the process of such generalization we do loose some valuable information but it allows us to see immediately if we are going in the right direction or not. Our 12 affective terms can be grouped into four categories (see Figure 6.1) forming four quadrants in terms of arousal and valence dimensions. By analyzing 25 songs we found that a song can have all four regions being equally active or it can have from 1 to 3 regions being dominant. If we go further we can separate 11 different categories based on which regions were dominant compared to the whole emotional space (see Figure 6.19).

We found five songs with only one dominant region. The songs "Nothing Else Matters" (Figure 6.18), "Come Away With Me" (Figure F.6), "Starlight" (Figure F.17) and "What I've Done" (Figure F.22) are predominately *Sad*, while the song "Colorblind" comes out as *Happy* (Figure F.5). *Happy* and *Sad* are basic emotions therefore it is no surprise that in the cases where there is only one dominant region it turns out to be one of these two. In fact in every single distribution that we observed we found either *Happy* or *Sad*, or both (see Figure 6.20).

Eight songs showed strong emotional correlations in two regions at the same time. This is also very possible since dividing emotions into regions does not mean that we are limited to only one region at a time. For more complex emotions is it natural to be combined from different, or sometimes even opposite, basic emotions (for example *happy* + *sad* = *nostalgic*). The song "Iris" came out as *Happy* and *Angry* (Figure F.10). Both of these quadrants are *active* which matches with the song's up-tempo yet light style. The songs "San Quentin" and "Now At Last" turned out as *Mellow* and *Sad* (both are *passive*) which also fit the overall mood of the songs.

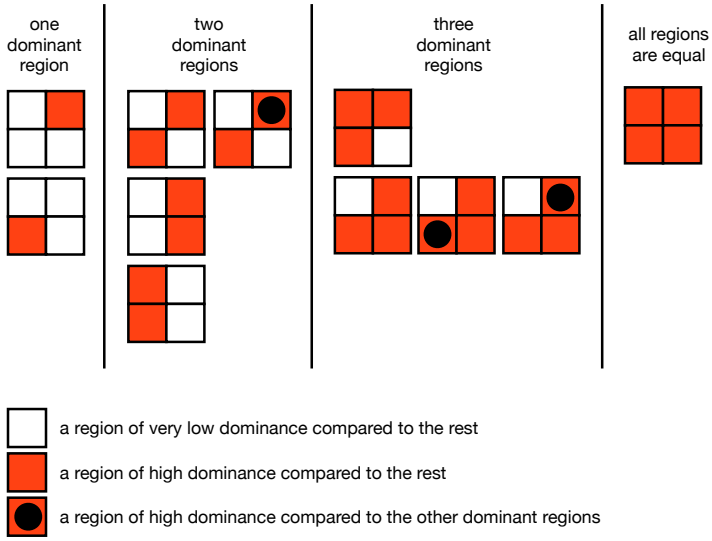


Figure 6.19: Different distributions of the accumulated correlations between song lyrics and 12 affective terms mapped into the *arousal–valence* quadrants.

"Time To Pretend" (Figure F.21), "Scientist" (Figure F.20), "Clocks" (Figure F.4), "Creep" (Figure F.7) and "The Pretender" (Figure F.19) are the examples of the songs where both *Happy* and *Sad* regions are dominant (the later two songs have the *Happy* region even more dominant compared to the Sad). As I mentioned earlier the combination of happy and sad can be interpreted as nostalgic, which would fit the mood for most of these songs. On the other hand TASA text corpus is not the best to separate these two emotions since their vectors appear to be very close to each other and come out as near neighbors. This can only be addressed by choosing the different text corpus, which we did in our current work (reflections on what difference does the new corpus make are presented in *Chapter 7* (see Figure 7.2)).

12 of the 25 songs have three or four regions being dominant. The interpretation becomes more complex but it still makes sense when we know what the songs are like. Nine songs have three out of four regions being dominant. In such case it may make more sense to talk about them "missing" one region instead of having three dominant ones. Six songs are "missing" the *Mellow* region, while three songs are "missing" *Angry* region. The first six songs can be further classified based on which one of the dominant regions stand out the most. "My Immortal" (Figure 7.5), "Everybody Hurts" (Figure F.8) and "Wonderwall" (Figure F.23) seem to be equally balanced in *Happy*, *Sad* and *Angry* regions. "Rehab" (Figure

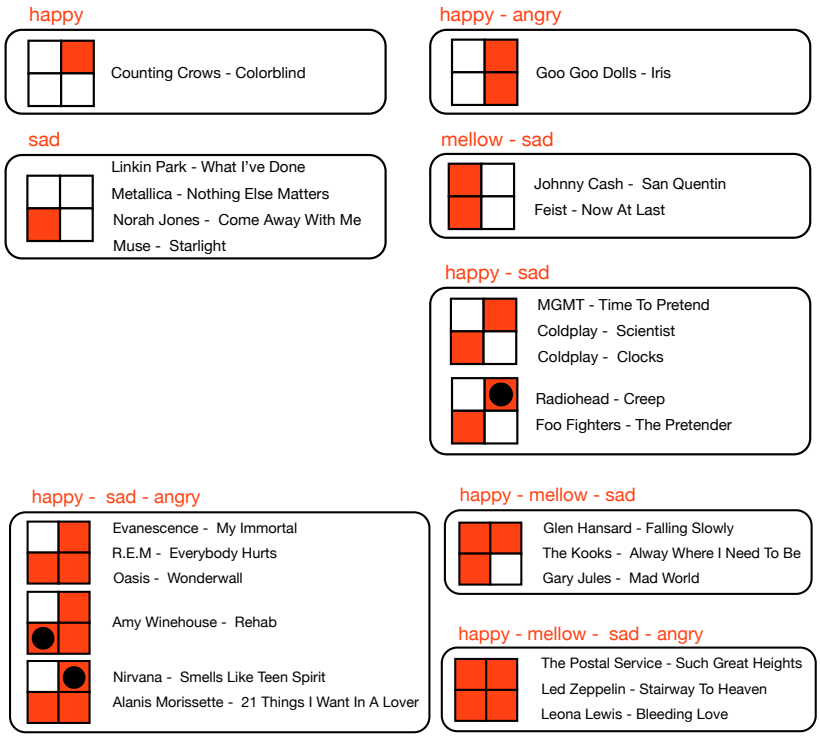


Figure 6.20: The 25 songs summarized into distributions based on emotional dimensions of *arousal* and *valence*.

F.13) has stronger presence in *Sad* while "Smells Like Teen Spirit" (Figure F.15) and "21 Things I want In A Lover" (Figure F.1) are stronger in *Happy*. Latter two make perfect sense in a way that they are both up-tempo songs.

"Always Where I Need To Be" (Figure F.2), "Falling Slowly" (Figure F.9) and "Mad World" (Figure F.11) are lacking the emotional presence in the *Angry* region. Last.fm listeners tag the first one as *happy* while the later two as *sad*. It matches our patterns on the general level but of course in order to tell more we need to look at the other details such as the pattern itself.

The last three songs seem to have very uniform emotional distribution once we generalize to the four regions. These songs are "Such Great Heights" (Figure F.18), "Stairway To Heaven" (Figure F.16) and "Bleeding Love" (Figure F.3).

Even though the bar-graphs with the accumulated correlation between then lyrics lines and the 12 affective terms allow us to highlight certain distributions

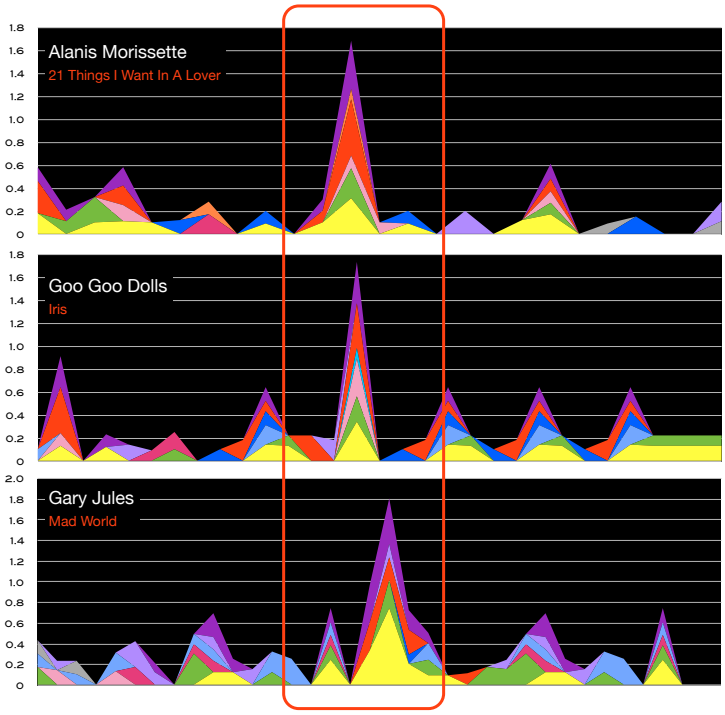


Figure 6.21: The sudden rise in emotional intensity in the middle of the song identifying *bridge*.

and thus group songs, we have to realize that this is just a fraction of the whole picture. Being projected into the vertical dimension such graphs represent the averaged values and ignore the fact that emotions in songs can evolve over time by forming changing patterns. Such information can be seen from the horizontal dimension (*time*) which is very important since the music by nature is very dynamic.

We did not perform the analysis of the songs emotional values in relation to time, but there are several points we noticed. Looking at the third graph (the one at the very bottom of the "song analysis" template) we can see that songs can be grouped by the number and distribution of peaks. In a number of the 25 songs we have one dominant peak – usually in the middle of the song. This corresponds to what in music is known as *the bridge* – the musically and lyrically different part of the song that usually follows after the second verse. It is interesting to see that it may be possible to notice the *bridge* as the peak of the emotional intensity of the lyrics alone (see Figure 6.21).

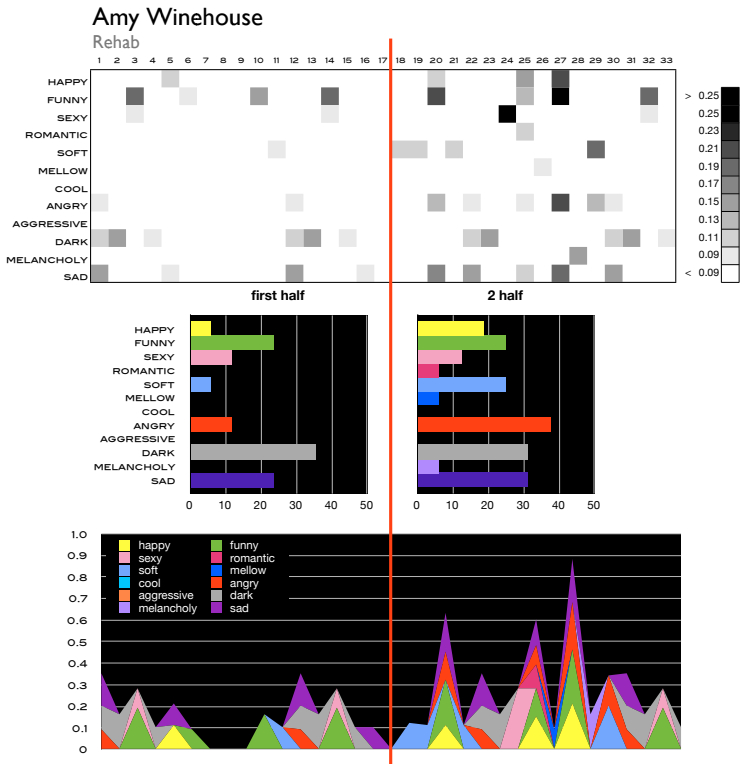


Figure 6.22: In the song "Rehab" the intensity raises in the second half of the song while the dominant emotions remain the similar.

We also noticed other types of changes in the pattern across the horizontal axis. That means that instead of summarizing the correlation values for the whole song we may identify certain parts of the song and make individual graphs for each part. In some cases (for example song "Rehab") the separate parts differ in term of overall intensity while keeping the same set of dominant emotions (see Figure 6.22). In the song "Creep" we can see three distinct peaks. The interesting part is that each of these peaks represent very similar emotions while the intensity level raises all the time (see Figure 6.23).

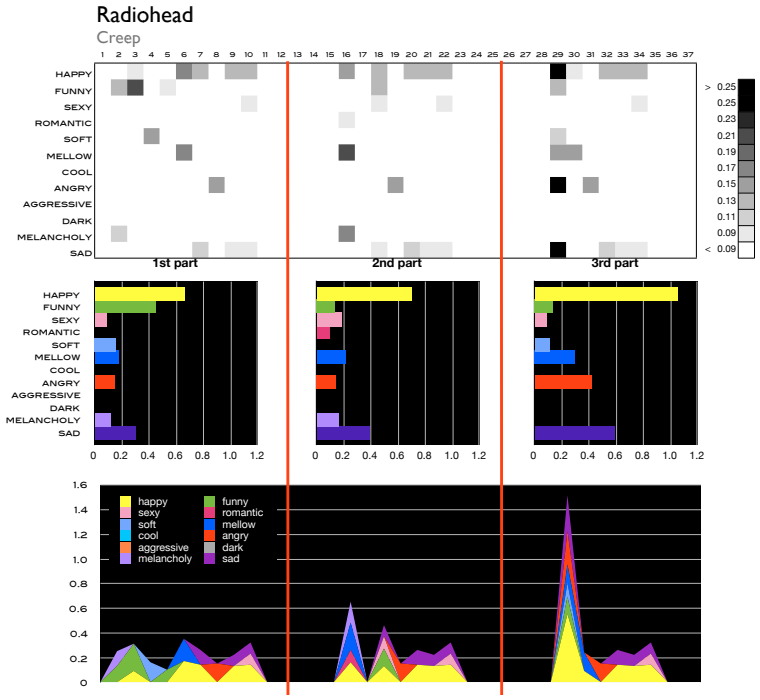


Figure 6.23: the the song "Creep" can be divided into three separate parts, every part reinforcing the previous one.

While in most of the songs we notice affective terms correlations remaining more or less of the same proportions compared to each other throughout the whole song (this can be seen from the first of the graphs in the *song template* where we do not see many lines crossing each other as the song progresses), in some cases we notice the emotional balance changing completely in different intervals in time. Such changes can be noticed in the songs "Such Great Heights" and "Now At Last". The song "Such Great Heights" could be divided in to three parts where the first one comes out more as *Funny* and overall very light; the second part of the song is more dominant in the *Angry* and *Sad* regions; finally the last part comes out as very *Mellow* and *Sad*. When we project the pattern into vertical axis we end up having the accumulated correlation values that trigger all four regions, but what we see from the horizontal projection is that different regions dominate in different sections of the song thus giving us a more more detailed picture of changing emotional patterns in the song (Figures 6.24.

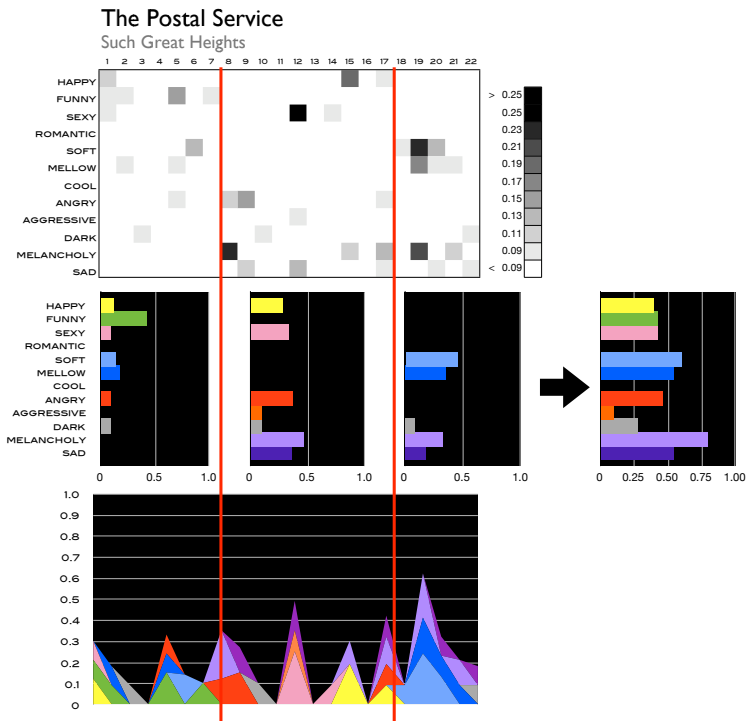


Figure 6.24: the the song "Such Great Heights" can be divided into three separate parts, every part reinforcing the previous one.

The song "Now At Last" (Figure 6.25) starts out very strongly triggering most of the affective terms at the same time – mostly *Happy* and *Sad*. But from the 9th line onwards into the song this changes – we get very constant triggering of the terms *Soft* and *Cool* both belonging to the *passive positive* region, also the term *Dark*. When we sum everything up we end up having having two dominant regions – *Mellow* and *Sad*.

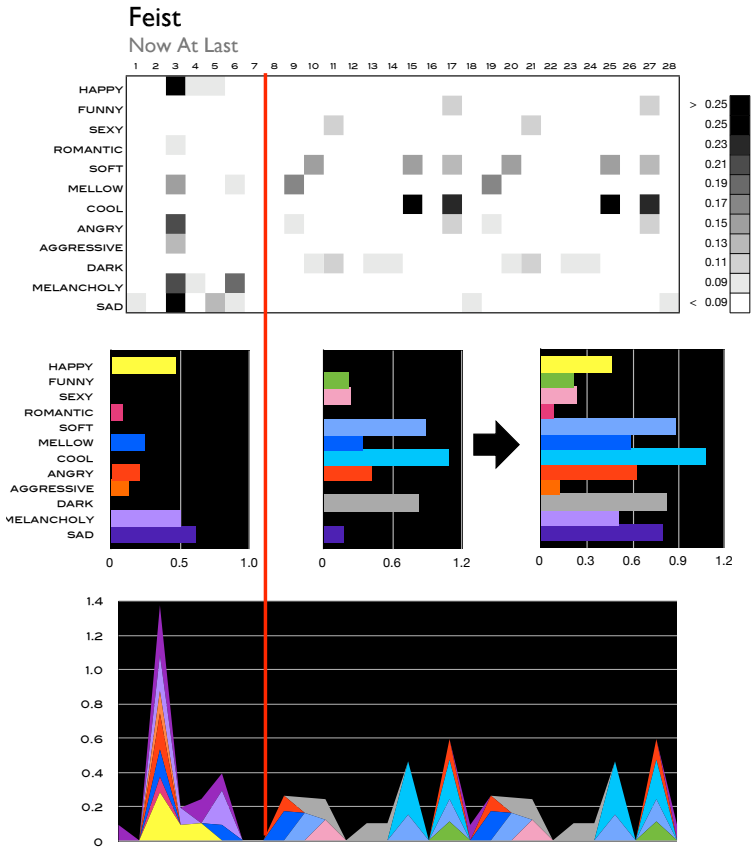


Figure 6.25: the the song "Now At Last" can be divided into three separate parts, every part reinforcing the previous one.

Further analysis is needed in order to map out the patterns of the songs in relation to the time axis. This is where we plan to put our focus in one of our upcoming papers. What we do see already is the amount of structure and that we can monitor changing contours of the songs by mapping out the correlations of their individual lines with the affective terms.

6.7 Validation of the results

So far I have presented a proposed methodology to extract emotional patterns from the TV programs or music songs by using the latent semantic information hidden in the TV synopsis or song lyrics. In both cases the emotional patterns are extracted from the text therefore a machine learning method was chosen to generate the emotional patterns. When it comes down to validation there are a number of elements in the methodology that have to be validated:

- Validation of the TV results against the TVA *Atmosphere* genres.
- Validation of the song results against the emotional tags from last.fm.
- Validation of the selection of the 12 affective terms (the top down part).
- Validation of the LSAs capability to understand the meaning of the words in english language (the bottom up part).

Let me start by validating the TV results and discussing the challenges that appear when we try to do such validation.

6.7.1 Validation of the TV programs vs. BBC genres

In first case we took a number of synopsis of the different TV shows and calculated their correlation with each of the 12 affective terms (see Figure 6.8). These were the results based on a single synopsis. In the case of "Eastenders" and "Two Pints" we also gathered five more synopsis for each program and performed the same type of correlation calculations (see Figure 6.12). How valid are our results and what should we validate them against?

All the BBC programs that we selected for analysis are labeled with genres from the five different *classification schemes* used by TV-Anytime (TVA) metadata specification. What we end up with are the combination of the affective terms with their correlation values for the given synopsis. The overall aim is to generate the emotional metadata automatically from the synopsis since we saw the lack of such metadata on the BBC side. There are a few programs that are labeled with *TVA Atmosphere* genres. In this case we could validate our results by comparing our graphs representing the correlation between the synopsis and each of the 12 terms (based on the LSA) with the emotional terms that these programs were labeled with by the BBC.

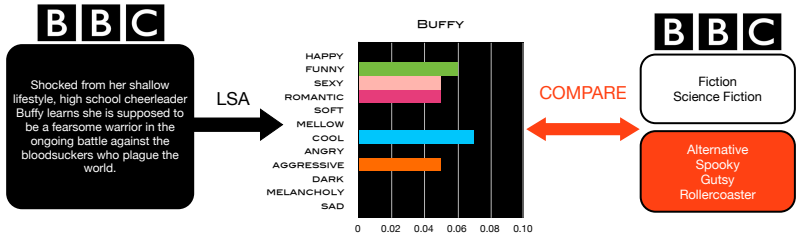


Figure 6.26: The validation of the extracted affective terms from the program "Buffy, The Vampire Slayer" against the labels that the program was given by the BBC.

Lets take look at the program "Buffy, The Vampire Slayer" as an example (Figure 6.26).

We can see that our extracted affective terms do not contradict the program synopsis, but when it comes down to validating them against the TVA genres we have several problems. The biggest problem is the expressiveness of the TVA *Atmosphere* genres – many of them are too vague to really show anything. From all the 53 *Atmosphere* genres only 21 have emotional meaning. For example, a genre *Alternative* is hard, if possible, to match against any emotional terms. Another problem is that different Atmosphere genres are on different levels – some of them show very basic emotions, for example *Happy* or *Sad*, while some show very complex emotions that are made by combining different emotional components, for example *Satirical*, *Breathtaking*, etc. This makes 1-to-1 mapping of the terms very complex.

Since the actual 53 words that TVA uses to express the atmosphere can not be directly mapped onto our 12 terms we can abstract both sets to a higher level where individual emotional words are grouped based on their similarity, which in turn is based on psychological experiments (for example the ANEW set (see Figure 6.1)). The methodology for the validation of the results could look like the one presented in the figure 6.27.

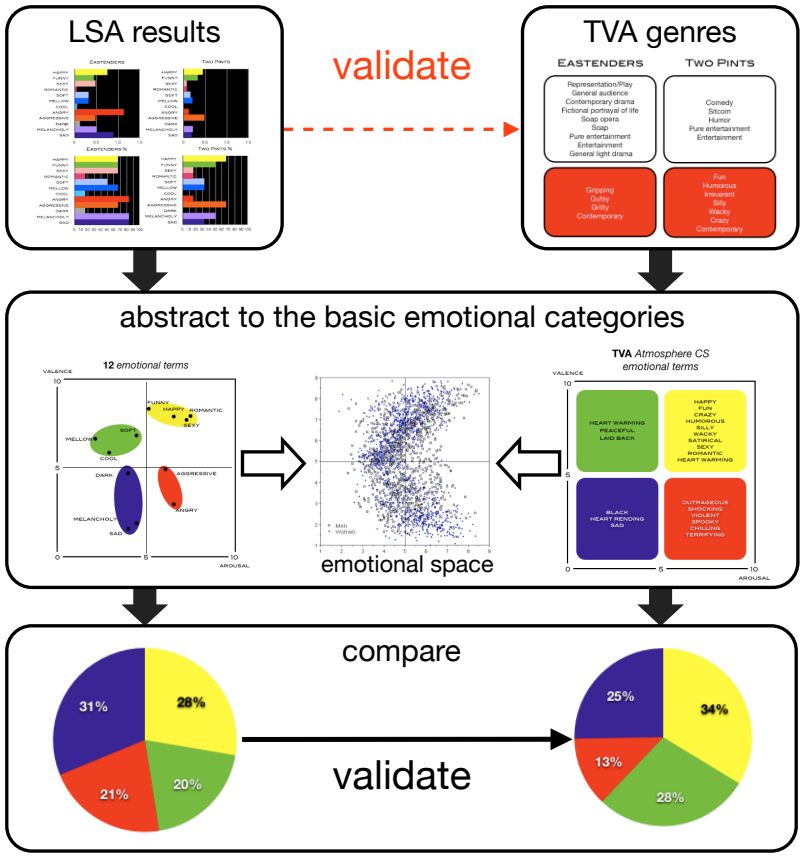


Figure 6.27: The methodology how to validate the emotional tags extracted automatically using LSA versus the BBC genres.

Since we are trying to validate our results by matching them to TVA emotional genres, lets take two TV programs that are well represented in terms of *Atmosphere* genres. I chose the programs "Eastenders" and "Two Pints" since both have been analyzed before in this thesis (*Chapter 5* and *Chapter 6*) (see Figure 6.28).

The biggest problem with the validation is the fact that we are trying to validate our affective terms against the TVA genres. We can question the validity of the genres because we know that they all come from the editors and may not reflect at all what the program is like. In the music case we also are trying to validate our results against the labels that media has, but in that case those labels are

generated by thousands of last.fm users and thus have, what we call, a *ground truth*. So instead of trying to validate our TV emotional patterns against the TVA Genres, we can instead see that the patterns reflect what we can normally expect from the genres like *soap* or *comedy*. In the case of "Two Pints", the correlations are generally smaller compared with the "Eastenders" which shows that a soap opera is heavier in terms of emotions compared with a comedy. "Two Pints" comes out as *Happy*, *Funny* and Aggressive – we may say that it matches the TVA *Atmosphere* genres very well. "Eastenders" are harder to match because of the vagueness of its TVA genres, but it does come out as more negatively charged compared with the comedy, which is expected from a soap opera.

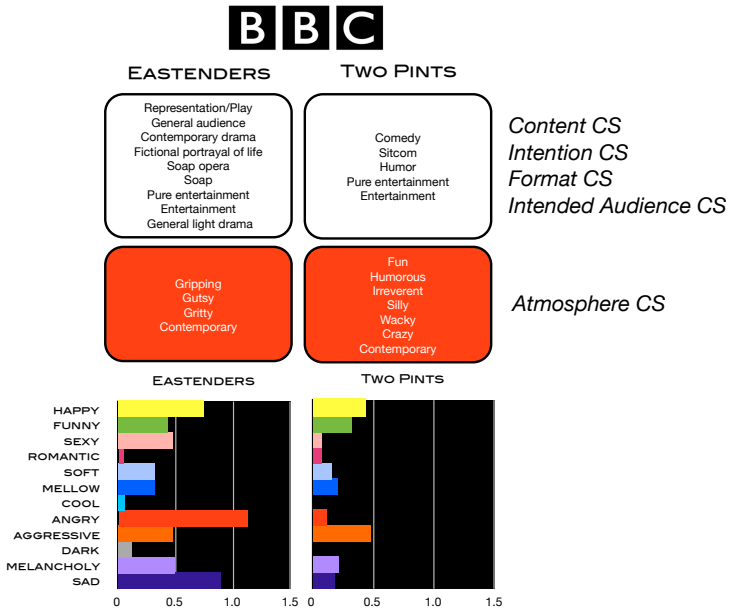


Figure 6.28: Both the TVA genres and the automatically generated emotional patterns for two of the BBC programs "Eastenders" and "Two Pints" presented next to each other for validation of the latter against the former.

6.7.2 Validation of the affective terms

What would happen if we took 20 or 50 emotional terms instead of the selected 12? In other words, how valid our 12 terms are. Even though the validation of why we selected these particular 12 terms comes from a number of differ-

ent psychological studies and experiments, we can still try out and see what would happen if we take more words. Our hypothesis was that expanding the number of emotional terms will result in similar patterns. To test that we took 21 genres from TVA *Atmosphere* and calculated the correlations between these genres and TV synopsis (we used the same two programs as before, only this time taking more episodes). The whole publication with the results [Petersen and Butkus, 2008a] is included in this thesis as an *Appendix* D. In here I would just like to show three figures explaining the results, since two of them are not in the original article.

In order to compare our 12 terms with the 21 genres we needed to abstract all of them into the categories based on the basic emotions. Even though there are many ways to group and classify emotions most of them agree on the small set of emotions being basic – *happy*, *sad*, *angry*, etc... We grouped our 12 terms into 4 groups, where each group could be represented by their prototypical emotion – *HAPPY*, *MELLOW*, *ANGRY* and *SAD*. After that we mapped our the TVA *Atmosphere* genres into these 4 categories (see Figure 6.29).

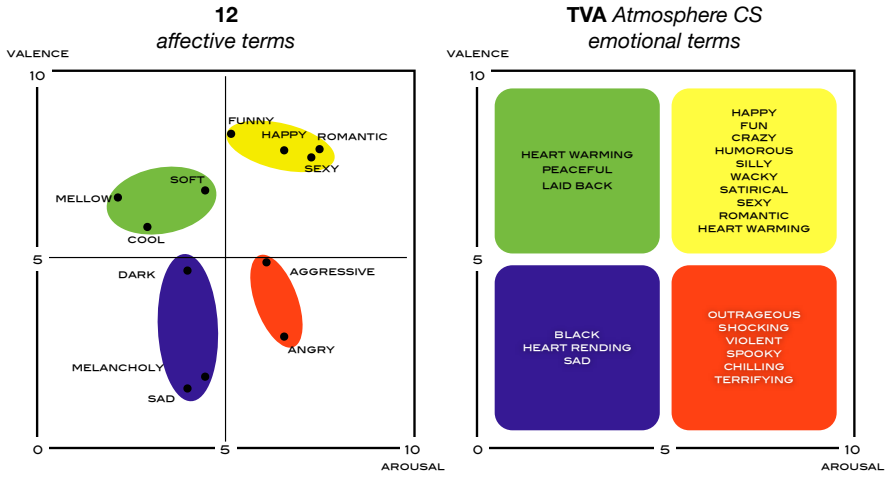


Figure 6.29: The mapping of the 12 affective terms and the 21 TVA Atmosphere genres into four categories based on the arousal and valence dimensions.

One we had the patterns from the LSA both for 12 and 21 terms, we plotted them as usually and then summed all the values in each row thus getting the accumulated correlation value of each of the 12 and 21 terms. Finally in order to compare the 12 terms with the 21 we mapped all of them into the four categories by averaging out their accumulated values (see Figure 6.30).

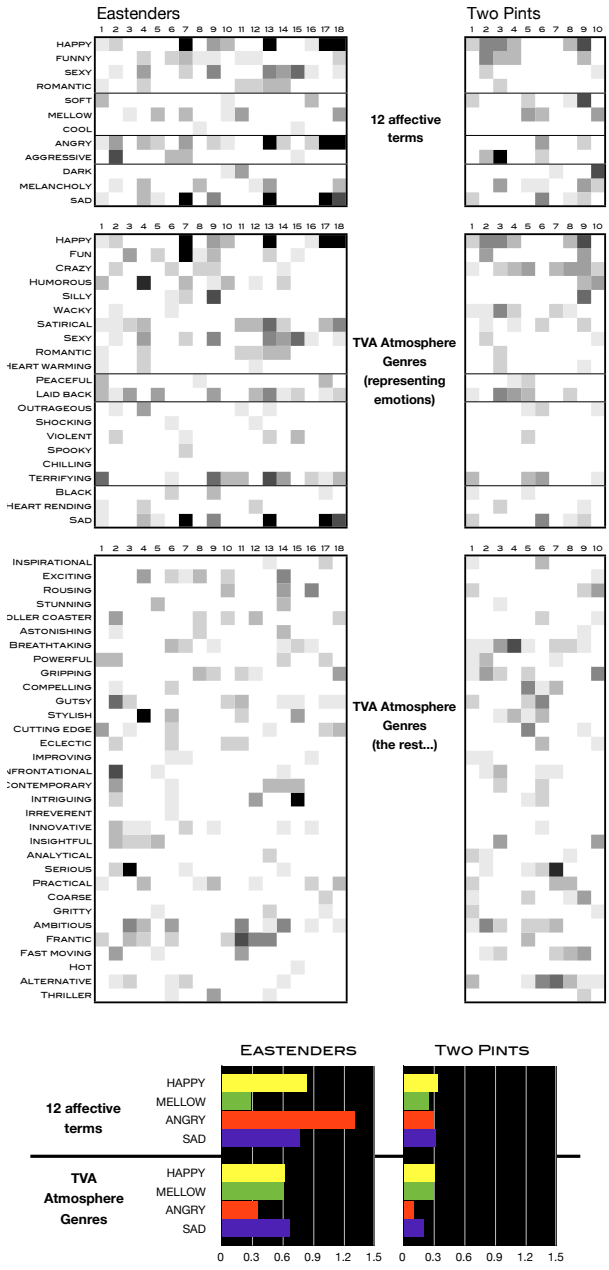


Figure 6.30: LSA correlation values of the 12 affective terms against the synopsis in comparison to the same correlations done on the 53 TVA Atmosphere genres.

What we saw was that the patterns and the overall emotional activity remains very similar if we take 12 or 21 terms. This can be credited to the LSA since all the correlations come from the bottom-up processing performed by the LSA unsupervised machine learning method. LSA is good at recognizing synonyms and overall words that are related. Therefore it is not surprising that both of the patterns turn out to be similar since a lot of the *Atmosphere* genres in TVA are perceived as synonyms even for humans (for example *fun*, *crazy* and *wacky*).

Such analysis shows that we need only a few words per every basic emotional category and that increasing the word number will not necessarily reveal more information. Once we have our basic emotions mapped out, we can derive more complex emotions by combining the 12 affective terms both vertically (which terms occur with which ones) and horizontally (how terms shift in time).

6.7.3 Validation of LSA for the language comprehension

LSA was chosen as a machine learning technique to calculate the correlations between the 12 affective terms (the top down part) and the textual input (synopsis or lyrics). The question is how valid it is to use LSA for such comparison.

The main reason by choosing LSA is because of the way it calculates the similarities between the words which are "...not simple contiguity frequencies, co-occurrence counts, or correlations in usage, but depend on a powerful mathematical analysis that is capable of correctly inferring much deeper relations" [Landauer and Dumais, 1997]. What we want from our chosen technique is for it to approximate how the human brain works when learning the relations between different words. In other words, one could say that LSA is a theory of the acquisition, induction, and representation of knowledge.

What we need to validate is how well does the LSA "understand" the english language. One of the standardized ways to evaluate how well non-english speakers understand english is by the TOEFL test. Part of the TOEFL test is the *synonym test* where participants are asked to choose the word from the list with the closest meaning to the given word. LSA was tested using a TOEFL test by Landauer and his colleagues [Landauer and Dumais, 1997, Landauer et al., 1998b]. The reported results showed that LSA achieved correct answers 64.4% of a time on a TOEFL synonym test, which turned out to be similar to the average score of non-english speaking participants [Landauer and Dumais, 1997].

For more discussion about how LSA simulates the *bottom-up* processes of our brain see the Appendix E where the latest article by the author and his colleagues is presented [Petersen et al., 2008].

6.7.4 Validation of the songs vs. last.fm tags

Some of the validation concerning the song results was already presented when grouping the songs into categories and relating them with the user generated emotional tagclouds retrieved from last.fm. Let me elaborate more on this topic and present more formal method to validate the results.

25 different songs were analyzed using the proposed methodology (Figure 6.15) in this thesis. How can we validate the results? First of all let me answer what can we validate the results against. Much like in the TV scenario we do have *labels* attached to every song. The main difference is the validity of the music labels versus the TV labels. In this thesis all the TV metadata was taken from the broadcaster BBC therefore it does not necessarily reflect what the media actually is but instead represents the labels used with intention to market and "sell" certain programs to the audience. In the music case such labels were taken from the last.fm social network where all of them were generated by thousands of users. One might ask how does that make music labels more meaningful compared to the TV labels. The way folksonomies work is that they allow users to label media using their own words where they are not restricted to any predefined categories. Once the users of such system start labeling media they can see the most popular labels (put by other users) for the media item. This encourages them to use the same words, given that they agree with what someone else has put already, instead of creating their own words. If we have enough users (thousands or even millions) then such process leaves us with the set of labels that most users agree on. This is meaningful because it gives us the *ground truth* – the shared meaning.

The presence of such ground truth is the main difference between the labels of songs in last.fm compared to the labels of TV programs at BBC. This is not about one media being music while the other is TV – it is all about the process of how the labels were acquired. The TV metadata from BBC could be compared to the *Web 1.0* approach (read-only type) whereas the user generated tags for the songs in last.fm social network is very much in line with the paradigm shift on the Internet identified as the Web 2.0 where users are just as much consuming as they are creating things.

The methodology to validate the results from analyzing songs is very similar to the one used for the BBC programs (Figure 6.31). We start with the correlation pattern calculated by the LSA and we need to validate it against the tags that last.fm user used to label that particular song. To be more precise I have to say that we have to compare our results with the *emotional* tags only since other popular tags would only add noise since they do not cary an emotional value (at least not the one that can be easily interpreted).

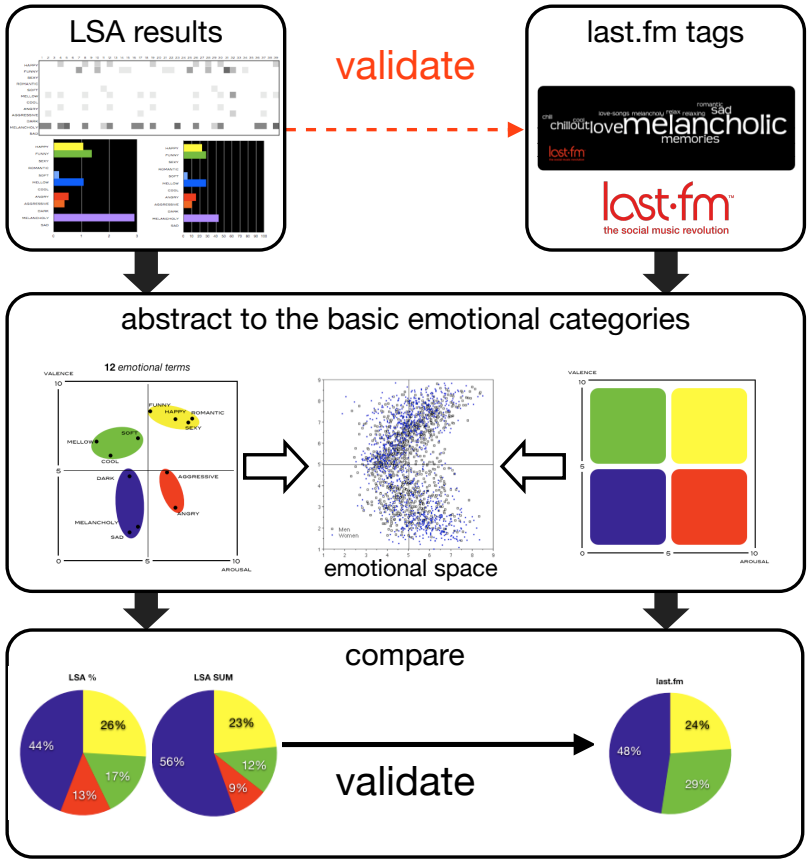


Figure 6.31: The methodology for the validation of the emotional patterns extracted from the lyrics using LSA versus the last.fm emotional tags.

When we chose the 12 affective terms we used last.fm as a ground truth to see what words people agree on when they express the emotional meaning of the songs. It may seen that because of that it will be easier to compare our 12 terms with the last.fm tags for a specific song. To some respect it is true – many of the user generated tags tend to be semantically close to one of the more basic emotions such as *happy*, *sad*, *mellow* or *angry*. This makes it easier to group last.fm tags into these four categories. On the other hand we still have tags representing more complex emotions such as *love*, *nostalgic*. We do not have such words as our affective terms therefore we can not do the direct mapping, also these emotions are combinations of more basic emotions which makes the mapping even more complex.

We can gather user generated tags from last.fm using their API which allows us to retrieve the top tags for every song in the form of an XML file. We can then take out the emotional tags from such files and list them together with their occurrence numbers. After that we just need to compare our bar-graphs with the ones that we got from last.fm. As I mentioned before, even though there is a big overlap of the exact same words used in both sets (ours and last.fm) last.fm tags also contain different words compared to our 12 terms. In order to compare the two sets we need to abstract each of them into more basic emotional groups (similar to what we did in the TV case). In the figure 6.32 you can see how such validation would look on a single song. To remain consistent I chose the same song as analyzed before – "Nothing Else Matters".

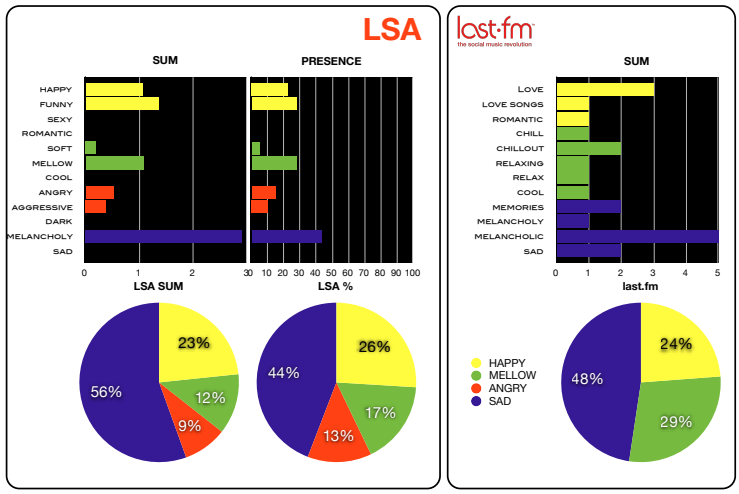


Figure 6.32: The validation of the emotional tags for the "Nothing Else Matters" song generated using LSA versus the last.fm tags for the same song.

Here I am comparing both the accumulated sum bar-graph (*SUM*) and the presence of emotion (*PRESENCE*) with the distribution of the user tags. We can clearly see that our results from LSA do not contradict what the listeners say about that particular song. All three graphs indicate the song as being mostly *Melancholy* (or in general belonging to the *Sad* region). I must highlight that here we are comparing the last.fm tags with only the vertical projection of the initial pattern that we retrieve from LSA. In the case of this particular song it seems that the vertical dimension seems to capture enough of the song meaning to match the last.fm tags, whereas in other cases it seems that the emotional distribution over time carries more meaning than their accumulated values. In either case we should take both of them in combination. The interpretation of

the time dimension is much harder to do since what we have on the user side are the tags reflecting the whole song instead of separate lines like we do in our calculations. Once we analyze much more data (we are building our song lyrics data set containing several hundred thousands of lyrics) we can then say much more about the interpretation of the patterns and thus to get more accurate validations against last.fm tags.

6.8 Conclusions

I started this chapter from the assumption that the emotional metadata is capable of identifying content items which might be perceived as similar and thus increase the number of relevant recommendations by capturing features across the traditional divide of categories. Such emotional metadata (TVA *Atmosphere CS* or emotional tags found in Last.fm), is not always available. The question then is – *Can we somehow extract similar emotional information automatically?*. This question leads to another one – *Where to look for such emotional information in the first place?*. Answering these two question was the main purpose of this chapter.

To answer the latter – it turns out that we can extract such emotional information from various unstructured metadata that is primarily targeted towards people and not machines. This includes all kinds of textual content descriptions, for example *synopsis*. Such information appears to be a good approximation of the inner structure of the media – its meaning, the way humans perceive it. While *synopsis* information is as far as we can go in the TV domain without getting into the actual audio-visual features of raw data, in music we can go a step further. This is because songs have one very special type of metadata – *lyrics*, which is both metadata and integral part of the content at the same time.

Now to get back the the first question - *how to extract emotions from synopsis and lyrics?* This chapter presented a novel methodology to do that. Theoretically the methodology builds on a number of different theories, while practically it is implemented using Latent Semantic Analysis as a machine learning method built, as the name suggests, to extract latent semantic meaning from unstructured text. Coupled with a set of affective terms serving as emotional markers, LSA was used to project our metadata (synopsis or lyrics) into a semantic emotional space, controlled by 12 selected affective terms. In other words, we projected our metadata into an emotional plane defined by psychological arousal and valence dimensions which provide the foundation for building conceptual spaces for media (based on emotional domain).

It was demonstrated in this chapter that such emotional information extraction is indeed possible, even considering how small the test sample was (8 TV programs and 25 songs) following by the proposed methodology for the validation of the results.

Chapter 7

Discussion and Conclusions

This last chapter presents the main points of the thesis. It starts with the discussion which focuses on the results and their interpretation. After that final conclusions are presented to overview what are the most important aspects that this thesis has touched upon, while also connecting all the different parts together. Next a number of scientific contributions are presented trying to keep the list of the actual contributions as clear as possible. This chapter ends by discussing several areas that can be still improved and falls under the *future research* category.

7.1 Discussion

The main idea of the thesis is that the key to the meaningful personalization of media is to combine the low-level features of the media itself (*bottom-up*) with our emotional responses (*top-down*). The notion of *similarity* is very central to both *top-down* and *bottom-up* parts. Since language can be used to encode both parts we can create a system where media is being evaluated using the language alone. From the implementation point of view it is also convenient since all the metadata that media carries is expressed as text and in most cases it is freely available.

The *top-down* part is modeled using different cognitive theories with the main focus on what we perceive as similar and how we categorize things. In this thesis the *top-down* element is materialized by selecting the 12 affective terms that are cognitively grounded. The *bottom-up* part focuses on how to automatically "understand" the meaning of media from the text which is based on building latent relationships between words. One of the machine learning methods – Latent Semantic Analysis – has been chosen for this purpose.

The convergence of the two parts happens when we project the 12 affective terms (*top-down*) into the semantic space of words where their meaning is expressed by the relations between different words (*bottom-up*). This process allows us to express the emotional value of media by patterns where we can monitor the correlation between our media and the affective terms evolving in time.

The proposed methodology was tested using a very small data set (8 TV programs and 25 songs) therefore it is too early for final conclusions. Nevertheless, the results show a lot of potential and confirm that we can indeed extract such emotional patterns from the text (the metadata). Discussion of the results and their interpretation was already presented in the *Chapter 6*. In this chapter I would like to highlight a few final points and present our latest research where we built our own LSA system and are using different text corpus compared to the one used in the previous chapter.

The empirical part of the thesis began by analyzing the TV programs using the traditional *genre* metadata gathered from the BBC, after that I turned towards another kind of information always present in the TV case – the *synopsis* which is different since it is unstructured and full of latent information to be extracted. Finally I ended up looking at the unstructured metadata present in music – the *lyrics*. One of the key points is to understand how these three types of metadata relate to one another and how they can be interpreted. There are a number of ways to look at this and I would like to discuss them here.

First of all we can look at those three types of metadata as having different abstraction levels (Figure 7.1). Imagine a piece of content, lets say a song, as the bottom level representing the actual data. Description in its very nature is related to extracting abstract information using certain rules. The key is to understand the difference between different abstraction levels and how the information changes when we move up through the layers of abstraction. The information that is the closest to the content is actually the content itself, since we are talking about a digital good to begin with. If we want to talk about the human readable metadata instead of going into the very bits of the raw content, then the closest that we can get is the information which is metadata while still being part of the content at the same time. Good examples of such metadata are the song lyrics, movie script, TV show subtitles, etc. That kind

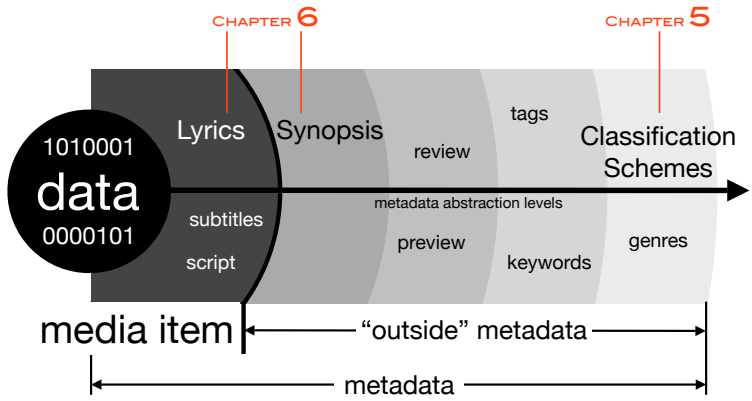


Figure 7.1: Abstraction levels for content metadata.

of metadata mimics the structure of the content as a whole, therefore it gives us the best approximation about what the content is like by preserving most of it's semantic structure.

If we move one step up, we end up being "outside" of the media, since the metadata and content are individual entities (one not being part of another). Examples here are all kinds of unstructured descriptive metadata that still attempts to describe an inner structure of the item. Here we can still draw connections between the structure of the metadata and the content, but since we are already on a higher level of abstraction, we can not get to the same level of detail. As a result our extracted structure becomes less accurate and much less fine-grained, on the other hand it seems easier to interpret because what we are processing has already been preprocessed by the person who wrote the description, e.g. synopsis, review, etc. This can be seen in the thesis if we compare the results from TV domain and music. TV synopsis is already abstracted to a certain level, while still trying to express the structure of the content, whereas song lyrics constitute part of the content and can be perceived as a raw material with the full amount of structure still in it. Of course I can not say that you can extract all the emotional context of the songs only from lyrics alone, audio features themselves play a very important role as well. What is clear though is that we can extract a fraction of the emotional structure that is there, and that fraction alone (being bigger or smaller depending on a case) can get us a long way. Getting back to the results, we can notice that emotional terms extracted from TV synopsis seems to be more easily interpretable and match the actual content well. On the other side, what we get from the lyrics is definitely more complex to interpret, has more noise in it and may seem not as precise after all. But this can mainly be interpreted as the fact that we are dealing with the

raw content and that it probably needs slightly different set of affective terms compared to the one that is used to process synopsis. Since affective terms serve as sensors we have to make sure they are calibrated to receive signals from the right level of abstraction.

Genres represent the highest level of abstraction, where we have the least amount of structure (genres themselves are very structured forming multilevel taxonomies, but they do not contain much of the structure about the content that they represent). Genres are already represented as prebuilt categories, whereas synopsis and lyrics are not. *Chapter 5* demonstrated the usage of TV genres forming categories, where the similarity was judged on the basis of how many features of one item overlap with another, and which genres are more informative that others (Shannon view). While in the *Chapter 6* the similarity was based on synopsis and lyrics can be judged on the basis of the inner structure and how different one item's structure is from another's (Kolmogorov Complexity view).

We can also look at these three kind of metadata – *genre*, *synopsis* and *lyrics* – as having different relation to *time*. Genres can be seen as completely static, they simply put things into categories. Synopsis can be seen as static as well if we take, for example synopsis for a movie or a TV show that is once stated and does not change. This can be referred to the eight BBC programs presented in the previous chapter, where they were described with a single synopsis. That resulted in much simpler, clearer interpretation, with probably much less potential compared with more dynamic cases. We saw some dynamic being added when several TV episodes were taken in a row positioning themselves along the time scale and thus representing the changing pattern of emotional balance over the course of several weeks – a sequence. In songs we saw the time window reduce to only a few seconds, as much as it takes for a single line of song lyrics to be processed by our brain. There we saw the increase of the complexity of the patterns.

When projected along their vertical axis such patterns can be summarized as overall emotional values of the song which allows us to group them based on which regions appear to be dominant (see Figure 6.19). If we take Conceptual Spaces point of view, then it is possible to look at these eleven patterns as an expression of relations between different properties. According to the the theory, this is precisely what moves us from just having a bundle of features to being able to express the concept. Knowing that the empirical data set was very small, I can not state that these are the final set of patterns. But once more empirical data can be projected into emotional space, then we may find out not only that we have more distinct patterns, but also what they represent, since we can draw similarities from the songs, once we have sufficient number of them.

The projection to the horizontal axis – the time – shows us how emotions build up over time and to monitor the overall emotional peaks and valleys of the song. We could suggest to use a "sliding window" several seconds wide (imagine this as 3 or 4 lines from the song lyrics) and see what kind of emotional contours we can find once we slide it over the song.

Another point I would like to highlight again is the relation between lyrics, music and emotions. In the previous chapter, before presenting the data on the music, I discussed the connection between the music and lyrics. According to numerous cognitive and neuroscience studies in various fields we can state that there exist a strong connection between the music and the lyrics. This is very important point of reference because it allows us to take only one of those two parts and still be able to draw conclusions about the actual media item, which, in case of the music, has both the music and lyrics. As it was pointed out, the music and lyrics are being processed in the similar and overlapping areas in our brain. Emotions here serve as a unifying medium when talking about music and lyrics, since there are big similarities in the way emotions are built up when we listen to music, compared to when we listen to speech. To some extent this can be illustrated by looking at the results presented in the last part of the thesis (the "music" section). The emotional tags that we extracted from the lyrics alone, seemed to very much reflect the general emotional balance of the whole song (music + lyrics) expressed by the thousands of users in the last.fm social network.

I also want to add a few words about some aspects on the interpretation of the results that we get using the suggested methodology. The main idea was that we use a selection of affective terms as markers in an emotional space, and we calculate cosine similarities between the text (synopsis or lyrics) and each one of the 12 affective terms. I discussed the reasoning behind the selection of the affective terms. What was not touched upon was the discussion what other factors influence the correlation and how can that be explained.

Here I would like again to use to Conceptual Spaces theory to explain what is going on. When we look at the user generated tags in the last.fm folksonomy, we can think of the tags as properties which come from the emotional domain of every person who is tagging reflecting the concept of the song that happened to be the object which is being tagged. The size of the emotional domain can be considered from two different perspectives. First of all, when deciding which emotional tag would fit best in a given situation, we draw upon our whole life experience, on the other hand we are able to understand words according to their context, in this case the content would be emotional value of music, meaning that when we say that the song can be annotated as, for example *dark*, we mean the emotional subspace, not the actual color. This is what we have on the user side resulting in the last.fm tags.

What happens on the "machine side" is that we rely on LSA to make correlations between the media and the affective term. LSA is purely automatic, this mean that it does not build on any external knowledge about the relationships between words, but instead it makes up its own knowledge about these relationships based only on analyzing big text corpus, where these word-word or word-document relationships are observed. This means that the two words will be as correlated as they appear in a given corpus. Therefore we do not have any absolute correlations, but every correlation has to be interpreted in relation to which text corpus was used. To some extent we could talk about general knowledge if we use text corpus that represents such general knowledge. This was exactly the case in this thesis, since I chose to rely on TASA text corpus which reflects general understanding and has shown good results in various studies. What is important here it to understand that since the corpus is general then the correlation between two words will also be drawn based on general knowledge, whereas in the "last.fm user" example we talk about the subspace of our general knowledge constituting to how we think of songs and emotions, and perhaps less influenced by our knowledge in other domains (for example, sports, physics, etc.). Therefore we can see one immediate source of errors coming from the usage of different domains on the user side(emotional musical domain) and system side (general domain) resulting in difference between which words come across as similar. The second source of errors comes from the fact that even though the corpus represents general knowledge, it may be not so broad on every single domain and definitely does not compare to a life-time of experience that the user has gathered in lets say the domain of music and emotions.

Why is this important and where do the errors show up? Let me illustrate with an example. The tricky part comes when our emotional words can be understood in other ways as well, for example we can talk about the emotional balance being *dark*, but generally we think of *dark* in physical terms as the absence of light. Here we have two different subspaces, and it is clear that the physical subspace is bigger compared to the emotional one – when we hear the word *dark*, our immediate reaction is to think of dark in physical terms, whereas if we know the context (if we happened to be talking about music) then upon hearing the same word *dark* we draw different conclusions since we are using another subspace. In the Norah Jones song "Come Away With Me" (Figure F.6) the affective term that is triggered most is *dark*, but if we look into the actual lyrics, we see that it was influenced mostly by the physical perception of the word rather than emotional. Another example can be a TV program "Flying Gardener" (see Figure 6.9) has a dominant affective term *cool* which comes from the connection with the the synopsis talking about *flying in the air* – again a physical domain.

This only stresses the importance of choosing the right text corpus, because it controls the context. Why did we choose TASA? Since to our knowledge there is

no sufficiently good and deep text corpus dealing with emotions only, we decided to go for very big and general corpus since we knew that it has been shown to work well in multiple domains. Latter we have built our own LSA setup which has several important advantages. First of all it allows us to create our own text corpus where we get to decide which texts go in it. Secondly it gives us the full control over the whole LSA mechanisms allowing to tune it in order to achieve better results (for example by choosing the different number of factors used in the SVD dimensionality reduction). In the following subsection I would like to present several findings from our latest work with our own LSA setup and different text corpus.

7.1.1 Using HAWIR text corpus

The moment we finished implementing our own LSA system the first thing we wanted to try out was the new text corpus and to see how is differs from the results achieved using the TASA text corpus. What we noticed before by looking at the emotional patterns of both TV synopsis and song lyrics is that the *Happy*, *Angry* and *Sad* emotions often came out coupled together. One way to explain this is to look at the nearest neighbors for each word. What we saw that for example the word *Happy* came out as the closest neighbor of the word *Sad*. This can be explained by the fact that these two words are indeed strongly correlated – not as similar but rather as the opposite. Happiness could be understood as the lack of sadness and vice versa. The question is do we want that type of correlation to be reflected in our patterns. Another question is how the patterns would be different if we use another text corpus where *Happy* and *Sad* would not come out as the closest neighbors.

To answer that we constructed our own text corpus consisting of 22,829 terms found in 67,351 paragraphs of poetry and literature from Harvard Classics, Wikipedia pages about music and news articles from Reuters (HAWIR).

In the figure 7.2 you can see the nearest neighbors of the word *Sad* for both TASA and HAWIR text corpora, while in the figure 7.3 you can see that HAWIR performs as well, if not better, than TASA in terms of the TOEFL synonym test (a standard procedure to test a non-english speaking person's knowledge of english language). From the 80 TOEFL questions HAWIR answered 71.25% correctly while the TASAs score was 64.4%. On the horizontal axis you can see the number of dimensions used when reducing the dimensionality of the *word-document* matrix. It could be compared with the figure 6.4 where it is showed that TASA seems to be working best with 300 dimensions while HAWIR seems

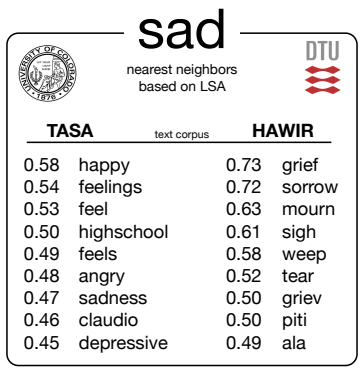


Figure 7.2: The nearest neighbors for the word *Sad* for both TASA and HAWIR text corpora.

to be optimal around 125 dimensions[1].

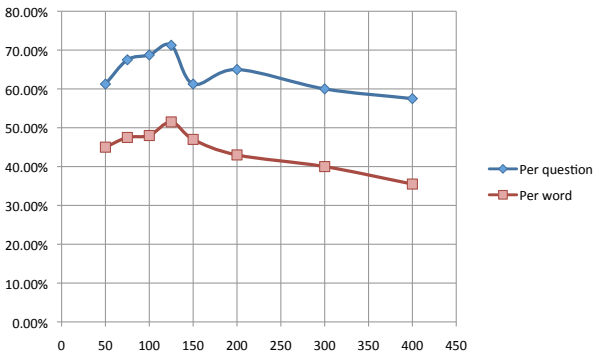


Figure 7.3: The percentage of the correct answers to TOEFL test based on HAWIR corpus with the horizontal axis showing the number of dimensions used in the reduction of the original matrix.

What exactly does the new text corpus add? Lets see by applying it on the TV synopsis and latter on the lyrics. In the figure 7.4 you can see the correlation patterns between the affective terms and the synopsis of the six consecutive episodes for the BBC TV shows "Eastenders" and "Two Pints". An interesting thing to notice is that the HAWIR corpus differentiates between *Happy* and *Funny*, while in the TASA these two emotions come out together (see Figure

[1]Such number of optimal dimensions can be only tested empirically

6.14). As human beings we know that there is a big difference between the feelings happy and funny. It fits the two programs well the fact that the comedy "Two Pints" comes out as *Funny* while the soap "Eastenders" appear to be more *Happy*. Such higher level separation of emotional components can be very valuable when trying to classify the media beyond the basic categories based on arousal and valence dimensions.

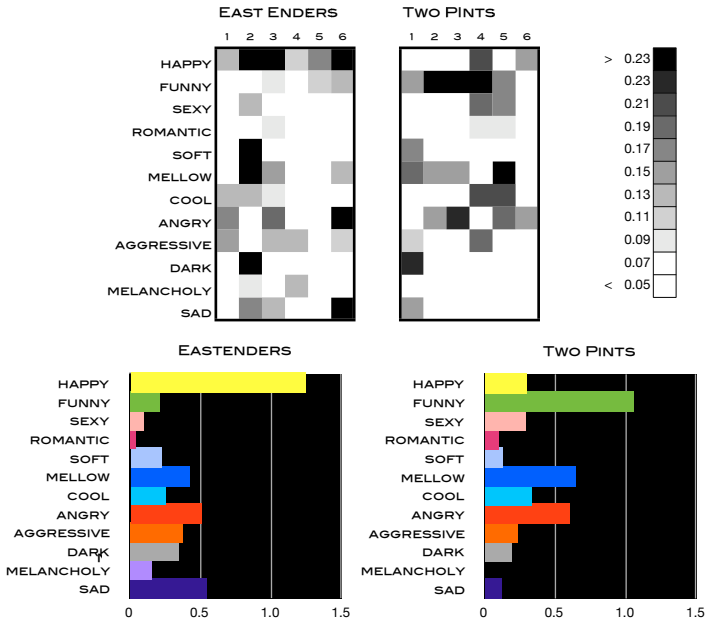


Figure 7.4: The emotional patterns of the TV shows "Eastenders" and "Two Pints" using HAWIR text corpus

We also tried out the new corpus on the lyrics. In the figure 7.5 you see the analysis of the Evanescence song "My Immortal" based on the TASA corpus while the figure 7.6 show the same song based on HAWIR. First of all we notice the different levels of saturation since in the HAWIR corpus affective terms on average come out as more correlated to the lyrics. What is the most interesting is what do they add up and how does that validate against the last.fm tagcloud compared to the TASA example of the same song.

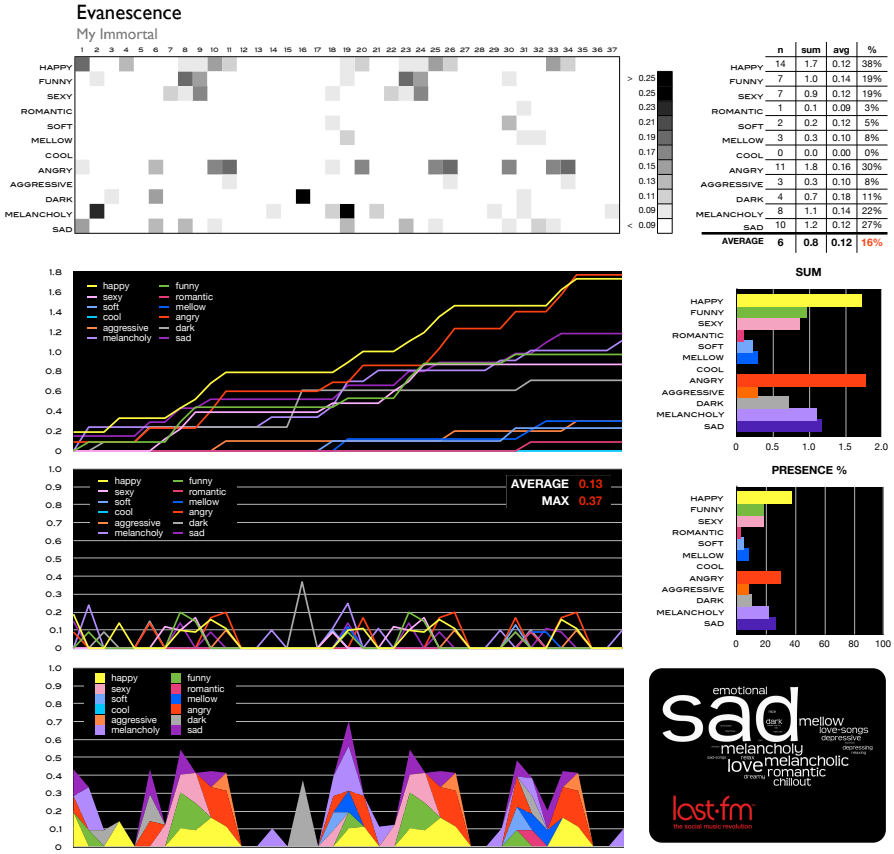


Figure 7.5: The emotional patterns of the Evanescence song "My Immortal" based on TASA text corpus.

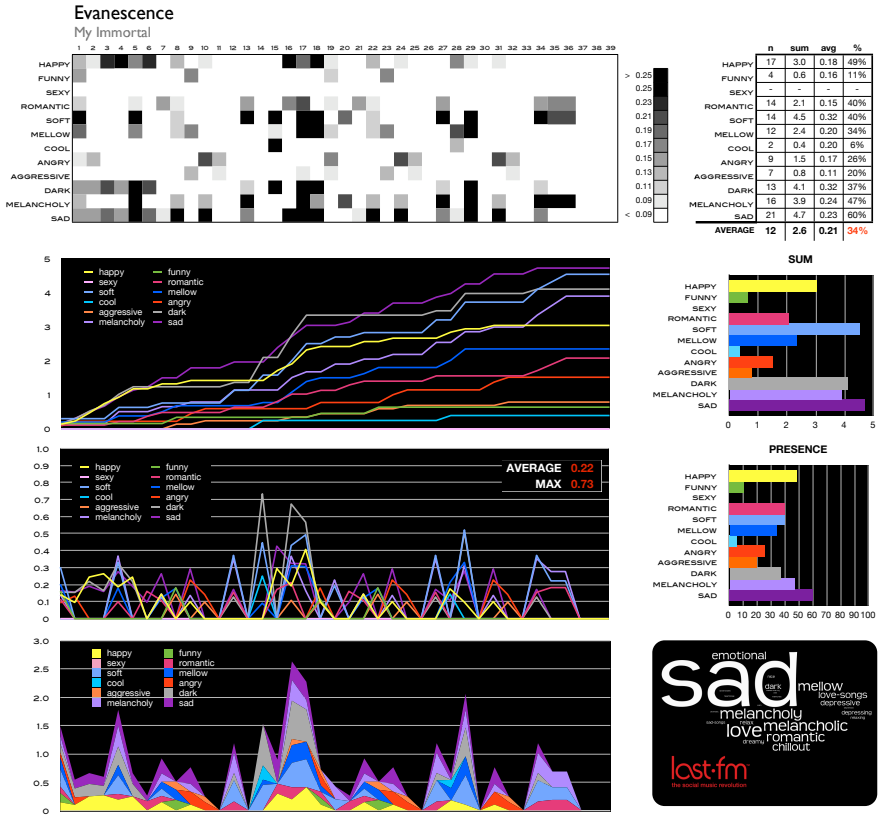


Figure 7.6: The emotional patterns of the Evanescence song “My Immortal” based on HAWIR text corpus.

We can see that the TASA version comes out as *Happy*, *Angry* and *Sad*, while the HAWIR labels this song as *Sad* and *Mellow* and *Happy*. If we generate the tagclouds based on the LSA analysis of the song for each text corpus individually and compare with the user-generated tagcloud we can see that the HAWIR tagcloud matches perfectly while TASA tagcloud comes out as different (see Figure 7.7).

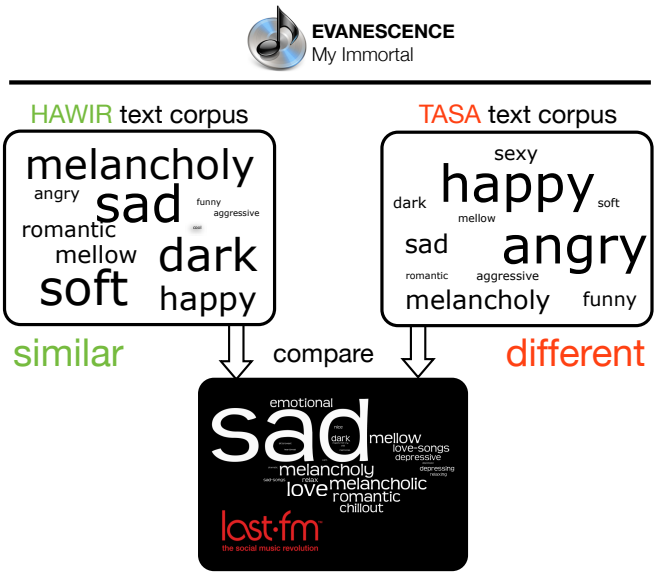


Figure 7.7: The validation of the emotional tags generated automatically by using LSA and both of the text corpora (HAWIR and TASA) versus the emotional tags from the last.fm social network.

I do not intend to claim that HAWIR text corpus is better compared with the TASA, instead I want to show that the selection of the text corpus has a major influence on the overall results. Once we analyze thousands of songs (we are in the process of analyzing 100,000 song lyrics using HAWIR) we can make stronger conclusions about the exact setup that is the most optimal for the proposed methodology to extract emotional patterns from the unstructured metadata.

7.2 Conclusions

In the course of this thesis I tried to look at the problem of media recommendations from a number of different perspectives, using a number of different theories and methodologies. Let me present my conclusions of what I found to be the most important aspects of these theories when it comes down to media recommendations.

Looking from the economical point of view, it can be concluded that for the last decade or so, we have witnessed the emergence of the new phenomena – the Long Tail of media. It has fundamentally changed the way we create and distribute media, and has opened up the gates into virtually unlimited selection of content. The main problem quite naturally turns out to be how we navigate and find things that interests us, leading to more intelligent recommendation systems.

What does it mean "intelligent" when we talk about the recommendations? The main requirement can be expressed as the ability to recommend relevant and novel media. In most cases relevancy comes down to the ability to find similarities between media items and then recommend the ones that are the most similar to either the user preferences, usage history, or the combination of the two. Different recommendation techniques were analyzed in the *Chapter 2* to present how they approach the similarity problem and where are the main bottlenecks. In collaborative filtering approaches sparsity seems to be the key problem since we can not effectively in real time process user-item interaction matrixes that contain millions of users and items. Content-based approaches have a few significant advantages, one of them is that it performs the most heavy calculations offline (thus increasing scalability), another one comes from taking the *item-item* similarities as the main criteria for recommendation, rather than just basing everything on user ratings and purchase history. While these two approaches are fundamentally different, they are often combined making up a hybrid recommendation systems. As one of the main bottlenecks that can be improved I identify the lack of ability to infer similarity between two media items. This made me to reduce the problem of media recommendation to the problem of similarity.

I started looking for an answer in the media description field, where I analyzed the different kinds of information that makes up the metadata and what kind of similarity knowledge each metadata type can provide. It appeared that the most potential lies in the unstructured metadata, which is primarily targeted to humans rather than machines. Another point is that emotional value appears to be very important when talking about video, and especially audio, content.

The search for the meaning of media has caused me to ask a question of, how do we perceive things, how do we know when things are similar. This led me to looking at the media recommendations from a cognitive perspective. Conceptual Spaces theory is applied in this thesis to analyze media in terms of its dimensions and knowledge domains, which in return defines properties and concepts.

One of the main hypothesis made at this point was that the emotional metadata is capable of identifying content items which might be perceived as similar and thus increase the number of relevant recommendations by capturing features across the traditional divide of categories. Such hypothesis was tested in chapter 5 where I presented the first empirical analysis of the TV personalization based on genre information. The first case showed the limitations of the approach to the similarity estimation based on the feature overlaps. But even as limited as it is, it has also highlighted the emotional genres being able to cross traditional categorization boundaries by trying to reflect the meaning of media.

Another hypothesis was that we can extract emotional information from various unstructured metadata since it reflects the inner structure of the media. I proposed a novel method to do so by combining the low-level features of the media itself (*bottom-up*) with our emotional responses (*top-down*). The method to extract emotional information from the unstructured metadata builds on the cognitive theories of knowledge representation and Latent Semantic Analysis simulating the way we learn language from the latent structure that it has by different words being related to each other through appearing in similar contexts.

Two separate cases were presented to illustrate the latent similarity knowledge extraction from the metadata. The first one demonstrated that such methodology is possible in the TV domain. Several BBC programs were selected to illustrate that it is possible to use TV program synopsis to extract emotional information form it, it was also shown that extracted emotions in most cases are a good indicator about the programs general atmosphere.

The second case showed how the methodology can be applied when analyzing songs. It was demonstrated that song lyrics can be projected into the emotional space one line at a time to create emotional patterns based on the triggering of the affective terms. As a result the emotional patterns were formed, and while at this point we can not explain fully what they mean, we see that such patterns exist, allowing us to talk about different songs as concepts which could lead to the creation of emotional music playlists. The method was validated by comparing LSA components of tracks to their corresponding tags taken from last.fm social network which represent a ground truth.

The final conclusion is that the early results have indicated, the extraction of the emotional components from the media is possible by following the proposed methodology.

7.3 Summary of contributions

The scientific contributions of the thesis can be grouped into two categories. The first group of contributions comes from looking at the problem of media personalization from three different angles – recommendation systems, media descriptions and cognitive science. Each of the three parts builds on the conclusions of the previous one (the first part builds on the introduction). It includes reviewing the state of the art, identifying the important elements and analyzing existing approaches. All these contributions give a foundation for the empirical analysis that is following after it.

- Analyze the problem of recommendation, by evaluating different approaches currently available in the market.
- Identify the metadata elements that are the most important for the media personalization.
- Present the cognitive science theories and models that represent the "human factor" (the way our brain works in term of categorization) in the media personalization system.

After the review of the theory follows the three main contributions. They are all based on empirical analysis of the data and builds on the three theoretical pillars described previously.

- Apply Gärdenfors' theory of Conceptual Spaces to modeling the concepts of media.
- Propose a model for constructing the emotional space, by selecting 12 terms to serve as emotional buoys or markers.
- Propose a novel approach for extracting emotional terms from either *synopsis* or *lyrics* metadata using Latent Semantic Analysis.

7.4 Future research

As it comes down to the future research, first of all more data need to be processed with the proposed methodology. This is being done at the moment by the author together with his colleagues (Michael Kai Petersen, Lars Kai Hansen and Martin Schwarts). So far we can witness that we can extract emotional patterns from analyzing unstructured metadata, but it is not possible to conclude yet what these patterns mean since there are still a number of unknown latent variables in the methodology.

Different elements of the methodology need to be fine-tuned to produce better results. The main component is the text corpus, which has been already switched from TASA to HAWIR. The text corpus is an essential element of the LSA which in turn represents the *bottom-up* part of the method by simulating the human learning of language.

Another element, which is equally important is the *top-down* part – the affective terms. Basing from the theory point of view we ended up having 12 affective terms. But once a large amount of data is gathered then we can use it to find out if our set of 12 terms is optimal. We also assumed that each of the affective term is weighted equally which may not be the case since we may have bias towards certain emotions.

Appendix A

Semantic Modeling Using TV-Anytime Genre

Published as a book chapter in *Interactive TV: A shared experience, Proceedings of EuroITV 2007 (ISBN: 978-3-540-72558-9), Springer-Verlag, Berlin 2007.*

Semantic Modelling Using TV-Anytime Genre Metadata

Andrius Butkus and Michael Petersen

Technical University of Denmark,
Center for Information and Communication Technologies,
Informatics and Mathematical Modelling,
Building 371, DK-2800 Lyngby, Denmark
{ab,mkp}@imm.dtu.dk
http://www.cict.dtu.dk

Abstract. The large amounts of TV, radio, games, music tracks or other IP based content becoming available in DVB-H mobile digital broadcast, offering more than 50 channels when adapted to the screen size of a handheld device, requires that the selection of media can be personalized according to user preferences. This paper presents an approach to model user preferences that could be used as a fundament for filtering content listed in the ESG electronic service guide, based on the TVA *TV-Anytime* metadata associated with the consumed content. The semantic modeling capabilities are assessed based on examples of BBC program listings using TVA classification schema vocabularies. Similarites between programs are identified using attributes from different knowledge domains, and the potential for increasing similarity knowledge through second level associations between terms belonging to separate TVA domain-specific vocabularies is demonstrated.

Keywords: personalization, user modeling, TV-Anytime, item similarity.

1 Introduction

The large amounts of TV, radio, games, music tracks or other IP based content becoming available in DVB-H mobile digital broadcast, offering more than 50 channels when adapted to the screen size of a handheld device, requires that the selection of media can be personalized according to user preferences. The TV-Anytime metadata architecture has been chosen as standard in DVB-H for description of content in the ESG electronic service guide [1], which similarly provides possibilities for user interaction or submitting preferences utilizing the 3G channel as return path. This paper presents an approach to build implicit user profiles based on the metadata associated with the consumed content by combining attributes from multiple TVA *TV-Anytime* controlled term vocabularies in parallel [2]. The data models forming the fundament for the TVA metadata rely on describing media, preferences or the usage environment based on predefined classification schema attributes for classifying e.g. the specific genre of a piece of content in terms of its category, format, atmosphere or intended audience. As the semantic description can

P. Cesar et al. (Eds.): EuroITV 2007, LNCS 4471, pp. 226–234, 2007.

be extended to capture different media features by combining attributes from different TVA knowledge domains, this paper will in the subsequent sections:

- Assess the semantic modelling capabilities of TVA classification schema attributes based on BBC program information sample data.
- Identify partial similarity between programs based on TVA genre attributes from different knowledge domains.
- Demonstrate the potential for increasing similarity knowledge through second level associations between terms belonging to separate TVA domain-specific vocabularies.

2 Related Work

Current research within personalization related to recommender systems often combine content based and collaborative filtering models as well as statistical knowledge discovery techniques. Whereas content-based filtering uses specific features of the media to produce suggestions for other items of a similar genre or starring the same actor, collaborative filtering recommends other items based on the preferences of users who have requested the same media using correlation or vector similarity. Systems providing suggestions of movies like the CinemaScreen film recommender agent [3] combines collaborative with content based filtering. It takes into consideration actors, directors or genres that have previously appeared in collaborative filtering results and thus uses the content similarity for recommendation of new items that have not yet been rated by other users. To further improve recommendations and compensate for a lack of overlap in items rated by different users, case-based reasoning [4] apply data mining of profiles to retrieve additional similarity knowledge by extracting frequently co-occurring items and define association rules between pieces of content that appear to share certain characteristics.

Whereas these techniques in hybrid combinations can be used to retrieve similarity knowledge of items and users, a perhaps even more critical aspect is the selection of the features, which characterize the content and thus serve as a fundament for defining similarity. In the TVA metadata architecture these features are controlled terms selected from domain specific vocabularies listed in classification schemas. In the *iFanzy* recommender system the proposed TVA features are implemented to define item similarity based on a set of preferred channels, as well as being used to build collaborative filtering based on usage history to match the user to a stereotype group of other users with the same interests and viewing behavior [5]. In another content-based approach the TVA metadata attributes have been assembled in a hierarchical user model mirroring a taxonomy of TV program genres reflecting the features of the consumed media [6]. As less emphasis seems to have been directed towards how the features may complement eachother, the aim of this paper is to assess the potential for increasing item similarity knowledge by implementing multiple TVA domain specific attributes in parallel and thus extend the semantic dimensionality beyond traditional content genre hierarchies.

3 Semantic Modelling

Similar to the original MPEG-7 concept, the TVA Phase 1 classification schemas are indexing tools using controlled terms for describing a particular aspect of the metadata associated with the content. If generating implicit user preferences based on the media that is being consumed, the TVA terms may thus be implemented to classify the content `Genre` along several dimensions simultaneously based on attributes belonging to separate domain-specific vocabularies like:

`Origination` e.g. cinema, studio, on location
`Atmosphere` e.g. crazy, exciting, sad, insightful, heart-warming, analytical
`Format` e.g. documentary, cartoon, play, hosted show, quiz contest, DJ, structured
`IntendedAudience` e.g. adults, single, children 0-3, professionals
`Content` e.g. news, finance, soap, fascism, poetry, grunge, sports
`Intention` e.g. pure entertainment, inform, advice, enrichment, education.

The broadcaster BBC has since 2005 made their digital TV and radio program listings available in TVA format [7]. Implementing a subset of the TVA metadata architecture the BBC program information is mainly constructed around a description of `Title`, free text `Synopsis`, `Keyword` listings and structured `Genre` information combining terms from the `Intention`, `Format`, `Content`, `IntendedAudience` and `Atmosphere` vocabularies.

Which vocabularies and how frequently they are used to generate the TVA `Genre` information varies according to the needs of the channel for adequately describing its content. All channels rely primarily on the `Content` taxonomy to categorize the `Genre` within sub categories like e.g. `soap opera`, `game show` or `daily news`. Such subcategory terms alone would often in conventional recommender systems constitute what makes up the concept of a `Genre` description, whereas in the TVA architecture this type of `Content` categorization is only one among several aspects. Programs on the channels CBBC and Cbeebies to a large extent implement `IntendedAudience` terms to define that the described `Content` categories are meant for different age groups of children. Channels like BBC Four, News 24 and Parliament differentiate their `Content` categorizations by adding terms from the `Format` vocabulary, providing labels like `documentary`, `cartoon` or `interview/debate/talkshow` in order to simultaneously describe the internal structure of the `Genre`.

The BBC main channels One, Two and Three, which offer a high diversity of programs with a mixed schedule of current affairs, drama and entertainment, in addition to the above classification schemas also include attributes from the `Atmosphere` vocabulary like `heart-warming`, `crazy` or `insightful` to capture emotional aspects which go across the conventional `Genre` catagorization of `Content`.

Together the attributes from the `Intention`, `Content` and TVA knowledge domains provide taxonomies of terms, consisting of sub category hierarchies up to four levels deep. As such the terms are mainly nouns, which narrow down classification to specific types of `Content` or sum up the structure of a media item in

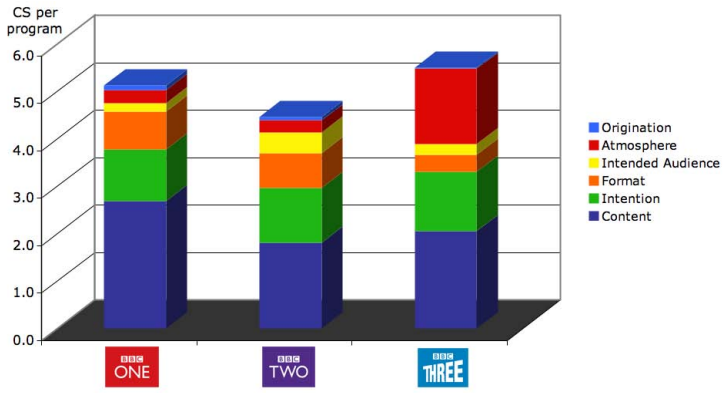


Fig. 1. Usage of TVA classification schemas for Genre description in BBC One, Two and Three program information over a twoweek period

regards to its `Format`. In essence the terms make up a top-down hierarchy for mapping numerous `Genre` features onto a small set of equivalence terms defined in the TVA classification schemas.

In contrast the attributes from the `Atmosphere` vocabulary are mainly adjectives capable of expressing associations, which instead of a hierarchy can be seen as spatially distributed. The distribution of the `Atmosphere` terms in itself might be more or less dense in regards to the number of adjectives available for describing the perceived responses when consuming the `Content`. Some of these terms define axes of opposites like `gripping` and `laid back`, or `happy` contrasted with `heart-rending`. The axes may intersect with planes of terms having an almost linear progression like `intriguing`, `astonishing` and `stunning` to emulate how compelling something is, or in the case of `gutsy`, `powerful`, `gritty`, `irreverent` and `confrontational` define degrees of radicalism. Yet other terms may appear isolated as dense sets of nuanced attributes like `humorous`, `fun`, `satirical`, `silly`, `wacky` or `crazy` capable of emphasizing specific aspects within the `Atmosphere`.

4 Item Similarity

Related to the aspects of entropy in information theory [8] a partitioning of items according to features reduces the data to a smaller number of significant characteristics and thus improves the effectiveness when predicting what media items to present in a personalized selection produced by a recommender system [9]. Seen in this light the TVA classification schemas provide a standardized selection of terms representing significant features of media items. So when considering information

entropy in relation to probabilities associated with selecting two similar programs among the available items, we might assume close to zero entropy if knowing that features fully describing two chosen items are identical. If knowing that not all of the features describing two items are identical we might assume only partial similarity with a corresponding increase in entropy.

Assuming that metadata attributes from the TVA classification schemas could provide sufficiently significant features for defining `Genre` item similarity, we have analyzed 2 weeks of BBC program description data. Assessing the average usage of different TVA vocabularies in BBC One, Two and Three (Fig.1) we first extracted data from BBC Three as it appeared to have roughly equal amounts of metadata attributes describing the `Genre` using controlled terms from both the `Content` and the `Atmosphere` knowledge domains (Fig.2).

BBC THREE	PR	CONTENT	INTENTION	FORMAT	INTENDED AUDIENCE	ATMOSPHERE	ORIGINATION	SUM
MONDAY 12	15	2.4	1.5	0.5	0.3	2.5	0.0	7.2
TUESDAY 13	16	1.9	1.3	0.4	0.2	1.1	0.0	4.9
WEDNESDAY 14	11	2.0	1.3	0.5	0.1	0.6	0.1	4.5
THURSDAY 15	15	2.0	1.2	0.3	0.3	1.7	0.0	5.5
FRIDAY 16	17	2.1	1.2	0.4	0.2	2.0	0.0	5.9
SATURDAY 17	17	2.0	1.1	0.1	0.2	1.2	0.1	4.7
SUNDAY 18	14	1.9	1.2	0.3	0.3	2.0	0.0	5.7
MONDAY 19	15	2.3	1.5	0.5	0.3	2.0	0.0	6.5
TUESDAY 20	17	1.9	1.2	0.3	0.2	1.1	0.0	4.7
WEDNESDAY 21	12	2.0	1.3	0.3	0.1	0.6	0.1	4.4
THURSDAY 22	14	2.3	1.6	0.4	0.4	2.3	0.0	6.9
FRIDAY 23	17	2.1	1.3	0.4	0.3	2.4	0.0	6.5
SATURDAY 24	16	2.1	1.1	0.2	0.3	1.8	0.1	5.4
SUNDAY 25	15	1.8	1.1	0.3	0.3	1.9	0.0	5.3
AVERAGE		2.0	1.3	0.3	0.2	1.6	0.0	5.5

Fig. 2. The usage of TVA Classification Schemas for Genre description in the BBC Three program information

Working from the hypothesis that the `Content` and `Atmosphere` vocabularies could be orthogonal we wished to analyze whether it would be be feasible to retrieve additional item similarity between programs by combining attributes from the two vocabularies. These programs would not necessarily be close in terms of `Content` categorization but might still be relevant for recommendation due to their overlap in `Atmosphere`. We therefore first analyzed to what degree the BBC Three programs could be seen as similar based on whether they would share one or more `Content` classification metadata attributes, and following whether also taking their `Atmosphere` descriptions into consideration would increase the number of perceived similar media items.

After that we extracted data from the BBC Two program information, which from the distribution of classification schemas seemed to suggest that the `Atmosphere` vocabulary was much less used when describing the `Genre` and that we consequently would expect little effect in terms of identifying additional item similarity. We here similarly first looked for overlaps between programs sharing on one or more `Content` classification terms, and following whether including `Atmosphere` would extend the selection with additional perceived similar items.

BBC TWO	PR	CONTENT	INTENTION	FORMAT	INTENDED AUDIENCE	ATMOSPHERE	ORIGINATION	SUM
MONDAY 12	40	1.9	1.1	0.8	0.4	0.4	0.1	4.5
TUESDAY 13	40	1.8	1.2	0.7	0.5	0.3	0.1	4.6
WEDNESDAY 14	39	1.9	1.2	0.8	0.5	0.3	0.0	4.8
THURSDAY 15	37	1.9	1.2	0.8	0.5	0.3	0.0	4.8
FRIDAY 16	38	1.9	1.3	0.9	0.5	0.4	0.1	5.0
SATURDAY 17	35	1.7	1.2	0.7	0.4	0.1	0.1	4.2
SUNDAY 18	25	1.6	0.9	0.4	0.2	0.0	0.2	3.3
MONDAY 19	42	1.9	1.2	0.7	0.4	0.3	0.0	4.6
TUESDAY 20	40	1.9	1.2	0.7	0.4	0.3	0.0	4.5
WEDNESDAY 21	38	2.1	1.1	0.7	0.4	0.3	0.0	4.7
THURSDAY 22	38	2.0	1.1	0.8	0.4	0.3	0.0	4.7
FRIDAY 23	40	2.0	1.3	0.8	0.5	0.5	0.0	4.9
SATURDAY 24	33	1.8	1.2	0.7	0.5	0.1	0.1	4.3
SUNDAY 25	26	1.5	0.9	0.4	0.2	0.0	0.2	3.3
AVERAGE		1.8	1.2	0.7	0.4	0.2	0.1	4.5

Fig. 3. The usage of TVA Classification Schemas for Genre description in the BBC Two program information

5 Results

Analyzing `Genre` item similarity in the BBC Three program information based on the `Content` attributes only, highlights some of the challenges content providers are facing when indexing media in a hierarchal structure. No less than 16 out of 28 programs overlap on the very top level of the taxonomy by being identically labeled `amusement/entertainment`. One level deeper in the `Content` taxonomy 9 out of the 16 programs are defined as `comedy` using a second attribute, while 4 and 3 programs are labeled as `non fiction/information` and `general light drama` respectively. The usage of `Content` terms is thus mainly concentrated within the upper layers of the taxonomy resulting in a relatively general classification. Fewer programs are defined based on the lower more detailed levels of the `Content` taxonomy resulting in very little overlap between items that are more accurately defined in terms of their `Genre`.

Secondly when analyzing the BBC Three program information on a program level and not just considering an average usage of classification schemas on a channel basis, it becomes evident that the `Atmosphere` terms are in reality only used for `Genre` description in 4 out of 28 programs, which can also be seen from the fluctuating distribution over the two weeks period (Fig.4). In this case among 2 out of the 4 programs the `Atmosphere` vocabulary terms can be seen as axes consisting of the terms `humorous`, `irreverent`, `satirical` and `silly`. Yet another program is defined along an axis of the terms `gripping`, `gritty` and `gutsy`. When associated with the actual programs in the BBC Three program information data this additional information does not extend the item similarity. These characteristics are already captured in the `Genre` description based on the `Content` classification, and as a result the number of identified similar programs is not increased when analyzing two weeks of BBC Three program information.

When going through the same steps of extracting items sharing one or more TVA attributes, instead analyzing BBC Two program information, a different pattern emerges. The use of attributes from the `Atmosphere` vocabulary is much less pronounced but more evenly distributed across 9 different types of programs.

232 A. Butkus and M. Petersen

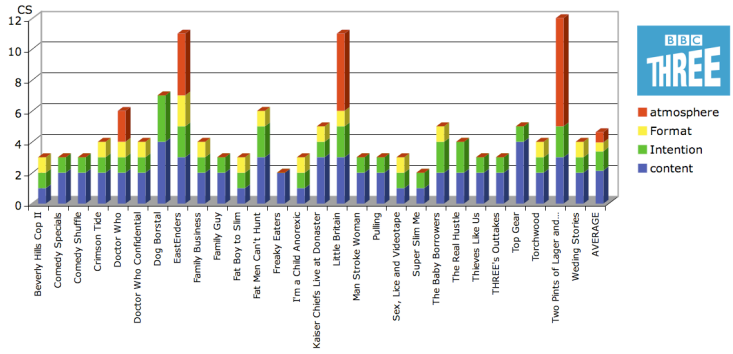


Fig. 4. Usage of TVA `Atmosphere` terms for Genre description in BBC Three program information related to specific programs during a twoweeks perio

Taking the auction show "Flog it!" as an example it would based on `Content` classification alone overlap with 33 other programs labeled as `amusement/entertainment`, while its more descriptive `fine arts` label from the lower levels of the classification taxonomy would not be shared by any other programs. Due to the elements of `consumer advise` and `quiz/contest` in the program it will further overlap with 2 and 6 other programs respectively within these more defined `Content` taxonomy sub-categories.

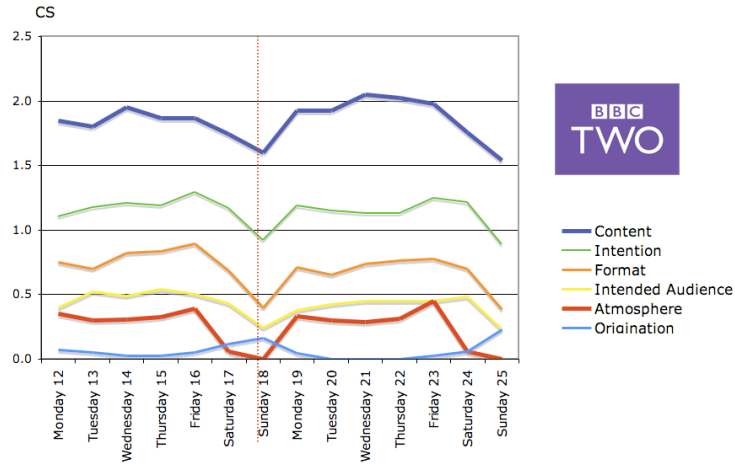


Fig. 5. Distribution of TVA `Atmosphere` terms for Genre description in BBC Two

Using attributes from the Atmosphere vocabulary "Flog it!" is also described in the program information as analytical, eclectic, insightful and astonishing. When filtering on the Atmosphere attributes it becomes apparent that this type of characterization makes the approch of the program rather than its fine arts subject stand out as the most significant feature. As a result one or two of these attributes characterizing "Flog it!" would be shared by 5 other programs, which are not closely related within the Content classification taxonomy. In this case it overlaps with programs defined by the following Atmosphere terms:

"Newsnight" - analytical, insightful, serious
"Newsnight Review" - analytical
"Gardeners World" - insightful, practical
"Escape to the Country" - analytical, practical
"TOTP2" - eclectic, rousing

None of these programs would be identified as similar to "Flog it!" if only taking Content classification into consideration when describing the Genre. If increasing our requirements when using Content classification and demanding that programs share at least 2 attributes based on differentiated terms indexed three levels down in the Content taxonomy, "Flog it!" would be identified as similar to only 8 programs. Adding terms from the Atmosphere vocabulary when analyzing item similarity of the BBC Two sample data would add 5 more programs, which could be considered relevant for recommendation when filtering media.

6 Conclusions

Though only very small samples of program information from the BBC channels Two and Three program information were analyzed, a number of issues related to retrieving item similarity between programs have been identified. Using only Content categorization as a basis for describing Genre will tend to identify similar items, which belong to closely related categories. One might argue that the terms belonging to the very attribute top levels of the Content taxonomy provide a too general categorization in order to efficiently identify similar program unless coupled with additional attributes from the more differentiated lower levels. At the same time this results in a scarcity of data due to lack of overlap between highly differentiated sub-categories of Content.

The Atmosphere attributes in the BBC Three sample data were associated with very few programs and the terms did not facillitate to further identify similar programs. Obviously the data sets were small but the results also highlight that the effectiveness of the Atmosphere attributes would depend on whether these terms are orthogonal to the description already captured by the Content classification. In the case of the analyzed program information from BBC Two the Atmosphere attributes were more evenly distributed among programs, and the potential for increasing item similarity between programs by combining the top-down Content classification approach with associative Atmosphere attribute terms was demonstrated.

References

1. DVB-H Mobile TV Implementation Guidelines: Nokia Profile of the Electronic Service Guide Datamodel for IP datacast over DVB, Release 1.5 (2006)
2. ETSI TS 102 822-3-1: “TV-Anytime; Part 3: Metadata; 1 Sub-part 1: Part 1 - Metadata schemas” (2006)
3. Salter and Antonopoulos: “CinemaScreen Recommender Agent: Combining Collaborative and Content-Based Filtering” vol.21 Issue 1 pp 35-41, IEEE Intelligent Systems, 2006
4. Wilson, Smyth & O’Sullivan: “Sparsity reduction in collaborative recommendation: A case-based approach”, International Journal of Pattern Recognition, Vol.17, No.5, 2003
5. Akkermans, Aroyo and Bellekens: “iFanzy: Personalised filtering using semantically enriched TV-Anytime content”, proceedings of ESWC, 2006
6. Pogacnik, Tasic, Meza and Kosir: “Personal content recommender based on a hierarchical user model for the selection of TV programmes” Vol.15, Issue 5, p.425-457 User Modeling and User-Adapted interaction, 2005
7. BBC Feeds & APIs, http://backstage.bbc.co.uk
8. Shannon: “Mathematical Theory of Communication”, Bell System Technical Journal, vol.27, p.379-423, 623-653, July, October, 1948
9. Sung Ho Ha: “Digital content recommender on the internet”, vol. 21 Issue 2 p.70-77, IEEE Intelligent Systems, 2006

Appendix B

Modeling Moods in BBC Programs Based on Emotional Context

Published as a book chapter in *Lecture Notes for Computer Science, Proceedings of "EuroITV 2008", Springer-Verlag, Berlin 2008.*

Modeling moods in BBC programs based on emotional context

Michael Kai Petersen and Andrius Butkus

Technical University of Denmark, DTU Informatics,
Building 321, DK.2800, Kgs.Lyngby, Denmark
{mkp,ab}@imm.dtu.dk
http://www.imm.dtu.dk

Abstract. The increasing amounts of streaming and downloadable media becoming available in converged digital broadcast and next generation mobile broadband networks will require intelligent interfaces capable of personalizing the selection of content according to user preferences and moods. We propose an approach to automatically generate atmosphere-like metadata from BBC synopsis descriptions, by applying LSA latent semantic analysis to define the degree of similarity between textual program descriptions and emotional tags in a semantic space.

Key words: personalization, emotions, LSA latent semantic analysis

1 Introduction

Since 2005 the broadcaster BBC has made their program listings available as XML formatted TVA TV-Anytime [1] metadata, which allows for describing media using complementary genre aspects, atmosphere as well as synopsis. We have in a related paper [2] analyzed how these metadata features may complement each other when applying more genre dimensions in parallel, and thus increase the number of relevant recommendations, by capturing similarities across the traditional divide of categories. In particular the TVA genre dimension atmosphere seemed able to identify programs that might be perceived as similar even though they belong to different genre categories. Extending this approach we propose in the present paper a method to automatically generate atmosphere-like metadata using the synopsis of TV programs. We outline in the following sections a framework for modeling emotional context using *last.fm* tags as markers in a semantic space, the methodology for extracting latent semantics, the retrieved results followed by a discussion of our early results based on BBC synopsis descriptions.

2 Affective terms

When investigating how unstructured metadata can be used to describe media, the social music network *last.fm* provides an interesting case. Despite the

idiosyncratic character of tags defined by hundred thousands of users, recent studies within music information retrieval have revealed that *last.fm* users often tend to agree on the emotional terms they apply to music. This correlation between social network tags and the specific music tracks they are describing, makes it possible to define high-level categories, which provide a simplified mood ground-truth reflecting the perceived emotional context of the music [3][4].

With point of departure in these findings we hypothesize that it might be possible to extract the emotional context of a TV program by projecting its synopsis into a semantic space, and use *last.fm* tags as affective buoys to define the textual description within emotional context. Drawing on psychological studies [5], establishing that emotional assessment can be reduced to a semantic differential spanned by the two primary dimensions of *valence* and *arousal*, we use these two axes to outline an emotional plane for a *last.fm* semantic tag space. The first of the these two dimensions describes how pleasant something is along an axis going from happy to sad, whereas the latter dimension captures the amount of involvement ranging from passive states like dark or soft to active aspects of excitation as reflected in tags like angry or sexy.

3 Latent semantics

As a machine learning technique which resembles cognitive comprehension of text, LSA *latent semantic analysis* [6][7][8] extracts meaning from texts by modeling the usage patterns of words in multiple documents and represent the terms and their contexts as vectors in a high-dimensional space. To retain only the most essential features the dimensionality of the original sparse matrix is reduced to around 300 dimensions. This reduced LSA space makes it possible to compute the semantic relatedness of synopsis and affective terms as the cosine of their vectors, with values towards 1 signifying degrees of similarity between the items and low values close to zero or negative signifying a random lack of correlation. In this semantic space a synopsis text and words which express the same meaning will thus be represented as vectors that are closely aligned, even if the terms are not literally co-occurring within the same context.

4 Results

Taking a selection of short BBC program descriptions as input, we compute the cosine similarities between a synopsis text vector and each of the selected *last.fm* emotional words. An analysis of the program "News night", based on the short description: *News in depth investigation and analysis of the stories behind the day(s) headline*, triggers the tags funny and sexy which might not immediately seem a fitting description, probably caused by these emotional terms being directly correlated with the occurrence of the words stories and news within the synopsis. The atmosphere of the lifestyle program "Ready Steady Cook!" might be somewhat better reflected in the synopsis: *Peter Davidson and Bill Ward*

challenge celebrity chefs to create mouth watering meals in minutes, which triggers the tag *romantic* as associated with meals. Another singular emotion can be retrieved from the documentary "I am boy anorexic", which based on the synopsis: *Documentary following three youngsters struggling to overcome their obsessive relationship with food as they recover inside a London clinic and then return to the outside world*, triggers the affective term dark. We find a broader emotional spectrum reflected in the lifestyle program "The flying gardener" described by the text: *The flying gardener Chris travels around by helicopter on a mission to find Britain(s) most inspirational gardens. He helps a Devon couple create a beautiful spring woodland garden. Chris visits impressive local gardens for ideas and reveals breathtaking views of Cornwall from the air.* The synopsis triggers a concentration of passive pleasant *valence* elements related to the words soft, mellow combined with happy. In this context also the tag cool comes out as it has a strong association to the word air contained in the synopsis, while the activation of the tag aggressive appears less explainable.

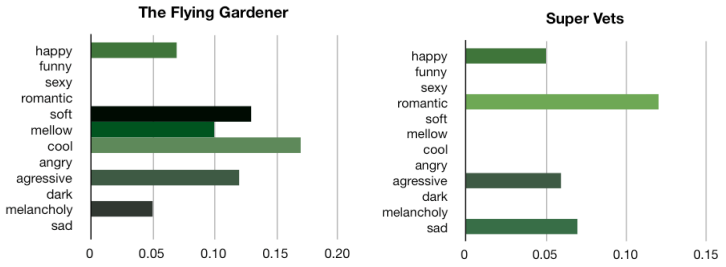


Fig. 1. LSA cosine similarity between the synopsis descriptions of "The flying gardener" and "Super Vets" against 12 frequently used *last.fm* affective terms

.

This cluster of pleasant elements is lacking in the LSA analysis of the program "Super Vets" which instead evokes a strong emotional contrast based on the text: *At the Royal Vet College Louis the dog needs emergency surgery after a life threatening bleed in his chest and the vets need to find out what is causing the cat (..)fits*, where both pleasant and unpleasant active terms like happy and sad stand out in combination with strong emotions reflected by the tag romantic. And as can be seen from programs like "The flying gardener" and "Super Vets" (Fig.1) the correlation between the synopsis and the chosen tags might often trigger both combinations of complementary elements as well as contrasting emotional components rather than a single monochrome feeling.

We proceeded to explore whether we could sum up a distinct pattern reflecting an emotional profile pertaining to a TV series, by accumulating the LSA values of correlation between synopsis texts and emotional tags over several

episodes. For this purpose we chose the soap "East Enders" and the comedy "Two pints of lager" and analyzed descriptions of six consecutive episodes from each series.

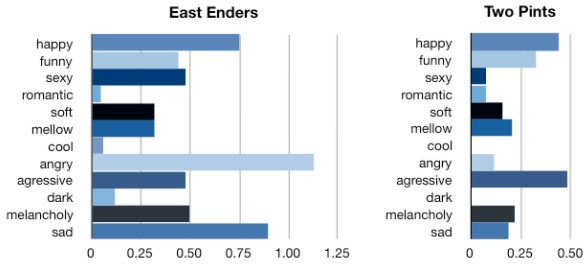


Fig. 2. LSA cosine similarity of the soap "East Enders" and the comedy "Two Pints" against 12 frequently used *last.fm* affective terms accumulated over six episodes

.

Even when only comparing the synopsis and emotional tags over six episodes (Fig.2), it appears that the accumulated LSA correlation values in the soap "East Enders" are roughly twice as high as in the comedy "Two pints of lager". The contributions of affective components in both histograms are unbalanced, but whereas the former series has a bottom-heavy emphasis on angry and sad emotions, the balance is reversed in the latter with a shift towards predominantly happy and funny elements complemented with soft and mellow aspects. These patterns can similarly be made out when considering the emotional components plotted over time for the soap and comedy respectively (Fig.3). The distribution in "East Enders" is much more dense and emotionally saturated reflecting aspects of *arousal*, while the character of "Two pints of lager" seems mirrored in a pronounced clustering of lighter elements of positive *valence* and an overall sparsity of excitation within the matrix.

5 Discussion

Projecting BBC synopsis descriptions into an LSA space using *last.fm* tags as emotional buoys, we have demonstrated an ability to extract patterns reflecting combinations of emotional components. Analyzing the emotional components reflected in the synopsis descriptions over a sequence of episodes, we have been able to separate these aspects into patterns defined by the sparsity and character of the distribution. While each synopsis triggers an individual emotional response related to a specic episode, general patterns still emerge when accumulating the LSA correlation between synopsis and emotional tags over consecutive

Fig. 3. LSA cosine similarity of the soap "East Enders" and the comedy "Two Pints" against 12 frequently used *last.fm* affective terms accumulated over six episodes

.

episodes, which enables us to differentiate between a comedy and a soap based on a textual description alone. We therefore propose that emotional components describing the content of media might be retrieved as latent semantics by using affective terms as sensors in a semantic space, and we suggest that LSA might be applied to extract structural patterns from synopsis descriptions as a basis for automatically generating mood-based recommendations. Though the synopsis descriptions trigger both combinations of complementary elements as well as contrasting emotional components rather than a monochrome affective response, they nevertheless pertain to distinct patterns which we speculate might be used as a basis to build emotional patterns capturing user preferences.

References

1. ETSI: TV-Anytime. Part 3. Metadata 1. Sub-part 1. Part 1 - Metadata schemas TS 102822-3-1,(2006)
2. Petersen, M., Butkus, A: Semantic modelling using TV-Anytime genre metadata In: Nagel,P. Cesar et al. (Eds.): EuroITV 2007, LNCS 4471, pp. 226–234, 2007. Springer, Heidelberg (2007)
3. Levy, M., Sandler, M.: A semantic space for music derived from social tags, Proceedings of the 8th International Conference on Music Information Retrieval, pp. 411–416, Austrian Computer Society (2007)
4. Hu, X., Bay, M., Downie, S.: Creating a simplied music mood classication groundtruth set, Proceedings of the 8th International Conference on Music Information Retrieval, pp. 309–310, Austrian Computer Society (2007)
5. Osgood, C.E, Suci, G.J., Tannenbaum, P.H.: The measurement of Meaning, University of Illinois Press, (1957)
6. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. Harshman, R.: Indexing by latent semantic analysis, Journal of the American Society for Information Science, Volume 41, p.391–407, (1990)
7. Landauer, T.K., Dumais, S.T.: A solution to Platos problem: The latent semantic analy- sis theory of acquisition, induction, and representation of knowledge, Psychological Review, Volume 104, p.211–240, (1997)
8. Dumais, S.T.: LSA and information retrieval: Back to basics, Handbook of latent semantic analysis, p.293–321, Erlbaum (2007)

Appendix C

Semantic Contours in Tracks Based on Emotional Tags

Published in the proceedings of *"Computer Music Modeling and Retrieval", Copenhagen 2008*

Semantic contours in tracks based on emotional tags

Michael Kai Petersen and Andrius Butkus

DTU Informatics
Technical University of Denmark, Building 321, 2800 Lyngby, Denmark
{mkp,ab}@imm.dtu.dk
http://www.imm.dtu.dk

Abstract. Despite the idiosyncratic character of tags applied to songs in social networks like *last.fm*, recent studies have revealed that users often tend to agree on the affective terms they attach to music. Using some of these frequently occurring words as emotional buoys to form a semantic plane of psychological valence and arousal dimensions, we project lyrics into this space and apply LSA latent semantic analysis to model the affective context of a number of songs. We compare the components retrieved from the lyrics with the user defined affective terms that constitute the tag clouds of the corresponding songs at *last.fm*, and discuss the potential for using LSA to extract structural patterns as a basis for automatically generating emotional playlists.

1 Introduction

Emotions in music are dynamically unfolding in time. Over the past half century these aspects of musical affect have been the focus of a wide field of research ranging from how emotions arise based on the underlying harmonic and rhythmical structures forming our expectations [1][2][3], to how we consciously experience these patterns empathetically as contours of tensions and release [4], in turn triggering physiological changes in heart rate or blood pressure as has been documented in numerous cognitive studies of music and emotions [5]. Recent studies suggest that musical structure to a larger extent than previously thought is being processed in "language" areas of the brain related to temporal structure and construction of meaning in general evolving over time [6]. Specifically related to songs both fMRI and ERP neuroimaging experiments point to linguistic and musical dimensions as being processed by similar overlapping brain areas, which seems to support the hypothesis that the linguistic and melodic components of songs are processed in interaction [7]. When retrieving songs from memory lyrics and melody appear to be recalled from two separate versions: one storing the melody and another containing only the text [8], while further priming experiments indicate that song memory is not organized in strict temporal order, but rather that text and tune intertwine based on reciprocal connections of higher order structures [9].

Despite the often idiosyncratic character of tags defined by hundred thousands of users in social networks like *last.fm*, a number of studies within the music information retrieval community indicate that users often tend to agree on the affective terms they attach to music, which be interpreted as a simplified mood ground-truth reflecting the perceived emotional context of the music [10]. We hypothesize that if music and text are cognitively processed in interaction, then a fraction of the tags attached to songs might possibly be retrieved by applying machine learning to extract latent semantics from the lyrics associated with the song. Selecting a number of previously identified frequently used emotional *last.fm* tags [11], as buoys to define a semantic plane of psychological valence and arousal dimensions, we project a number of song lyrics into this space and apply LSA latent semantic analysis [12], to model the correlation of texts and affective terms as vectors reflecting the emotional context of the songs.

Taking a largely qualitative approach we outline in the following sections: the methodology used for extracting latent semantics, the retrieved results and conclude with a discussion formulating our hypothesis.

2 Emotional tag space

The affective terms which are frequently chosen as tags by *last.fm* users form clusters around primary moods like happy, mellow and angry. Drawing on Osgood, Suci and Tannenbaums earlier findings establishing that emotional assessment can be expressed by the primary dimensions of *valence* and *arousal*[13], we use these two axes to outline an emotional plane for a *last.fm* semantic tag space. The first of these two dimensions describes how pleasant something is along an axis going from happy to sad, whereas the latter dimension captures the amount of involvement ranging from passive states like dark or soft to active aspects of excitation as reflected in tags like angry or sexy. We are applying twelve frequently occurring *last.fm* tags:

happy, funny, sexy
romantic, soft, mellow
cool, angry, aggressive
dark, melancholy, sad

These tags thus function as markers distributed across a semantic plane which we will need in the following for assessing the emotional context of the song lyrics and see whether we can retrieve part of the user defined tag-clouds associated with songs at *last.fm* based on the lyrics alone.

As a machine learning technique which resembles cognitive comprehension of text, LSA extracts meaning from paragraphs by modeling the usage patterns of words in multiple documents and represent the terms and their contexts as vectors in a high-dimensional space. The frequency at which terms appear and the phrases wherein they occur are defined in a matrix with rows made up of words and columns of documents. Many of the cells made up by rows and

columns contain only zeroes, so in order to retain only the most essential features the dimensionality of the original sparse matrix is reduced using SVD *singular value decomposition* to around 300 dimensions. This makes it possible to model the semantic relatedness of paragraphs and terms as vectors, with values towards 1 signifying degrees of similarity between the items and low or minus values typically around 0.02 signifying a random lack of correlation. In this semantic space paragraphs or words which express the same meaning will be represented as vectors that are closely aligned, even if they they do not literally share any terms. Instead these terms may co-occur in other documents describing the same topic, and when reducing the dimensionality of the original matrix using SVD the relative strength of these associations can be represented as the cosine or dot product of the vectors. The foundation for learning the associations between the synopsis paragraph and emotional words vectors are based on a large collection of documents, the frequently implemented standard *TASA* text corpus consisting of the 92409 words found in 37651 texts, novels, news articles and other general reading material that American students are exposed to up to the level of their 1st year in college.

3 Results

Taking the lyrics of twenty songs as input, we compute the cosine values between vectors representing each of the individual lines constituting the lyrics of a given song against each of the twelve selected *last.fm* tags in the LSA space and discard cosine values of correlation between lyrics and tags below a threshold of 0.09.

3.1 Accumulated distribution of emotional components

Analyzing a selection of songs the summed up values of LSA correlation between the lyrics and the *last.fm* affective terms appears to divide the data into roughly four groups:

Unbalanced sparse distribution reflected in a biased contribution of emotional components shifting the overall balance predominantly either towards happy or sad aspects, as in the songs “Come away with me” ,“San Quentin”, “What i’ve done”, “Always where i need to be”, “The pretender”, “Wonderwall” and “Starlight” (Fig.1). *Centered* distribution of emotional components shifting the emphasis towards the middle soft or mellow aspects with relatively less contribution from the outer extremes of happy or sad, as in the songs “Falling slowly”, “Now at last”‘and “Stairway to heaven” (Fig.1) *Uniform* distribution of components divided over the entire emotional spectrum as in the songs “Such great heights” ,“Mad world” ,“Bleeding love”, “Smells like teen spirit” and “Rehab” (Fig.2). *Balanced sparse* distribution combining contributions from the outer extremes of happy or sad with relatively less contribution from more central aspects like mellow and soft, as in the songs “Time to pretend”, “Nothing else matters”, “21 Things i want in a lover”, “Creep” and “Clocks” (Fig.2)

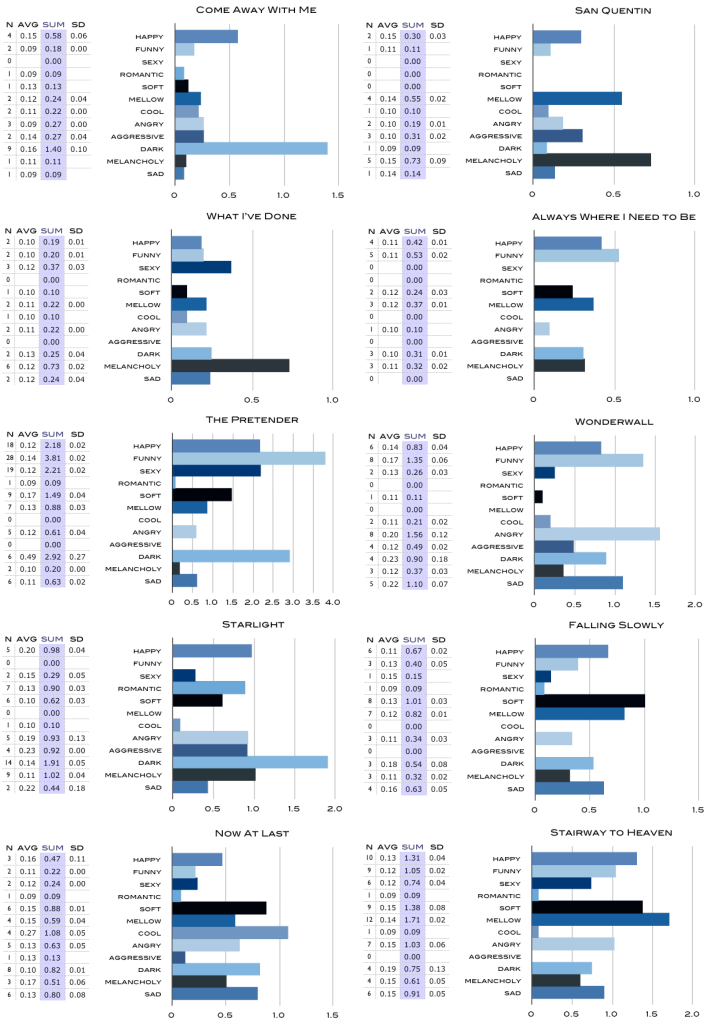


Fig. 1. Accumulated LSA cosine similarity between affective terms and the lyrics of "Come away with me" ,"San Quentin", "What i've done", "Always where i need to be", "The pretender", "Wonderwall", "Starlight","Falling slowly","Now at last", "Stairway to heaven"

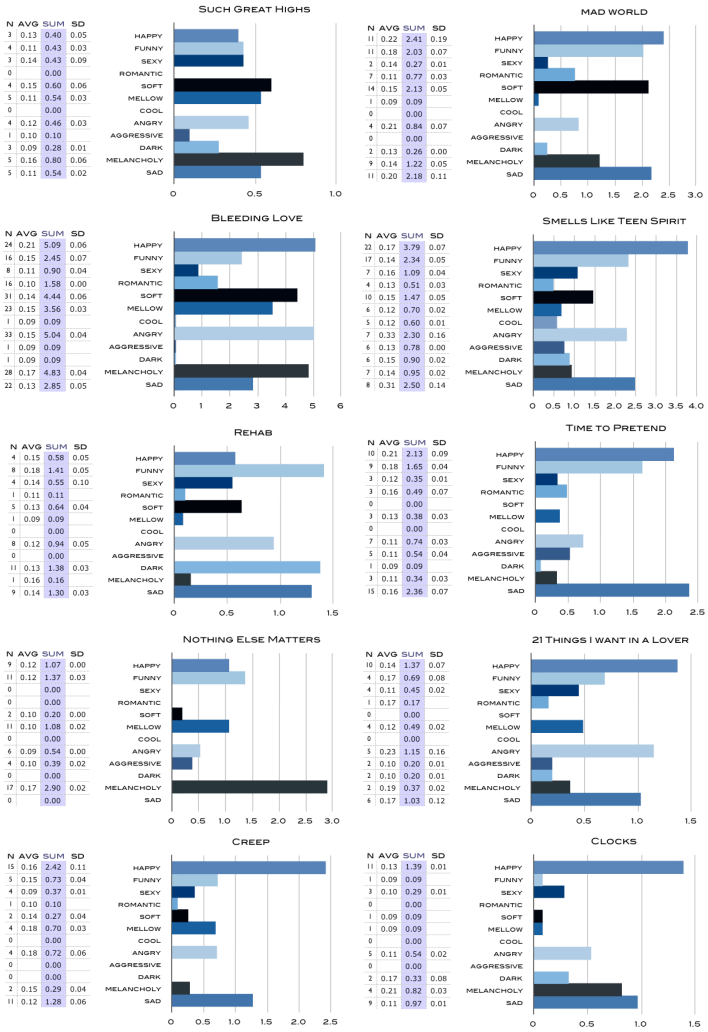


Fig. 2. Accumulated LSA cosine similarity between affective terms and the lyrics of "Such great heights", "Mad world" ,"Bleeding love", "Smells like teen spirit", "Rehab", "Time to pretend", "Nothing else matters", "21 Things i want in a lover", "Creep", "Clocks"

3.2 Distribution of emotional components over time

Plotting the LSA values of the individual lines in the lyrics of the entire song over time, provides a view of the distribution of emotional components which mirrors the structure of patterns and changing tension in the song along the horizontal axis. Vertically the color groupings indicate which of the aspects of valence and arousal are triggered by the lyrics as well as their general distribution in relation to each other. Any color will signify an activation beyond the cosine similarity threshold level of 0.09, and the amount of saturation from light to dark signifies the degree of correlation between the song lyrics and each of the affective terms.

Fig. 3. Color coding of LSA cosine similarity above the threshold value of 0.09, where the amount of saturation from light to dark signifies increasing degree of correlation between the song lyrics and each of the affective terms

The contribution of each emotional component apparent in the overall LSA values of the lyrics, can be made out when considering their distribution as single pixels over time triggered by the individual lines in each of the songs. Analyzing which components are predominant and their overall contribution in the lyrics, the LSA plots can roughly be grouped into four categories of *unbalanced sparse*, *centered*, *uniform* and *balanced sparse* distributions:

Fig. 4. LSA cosine similarity between the lyrics of "Come away with me" and 12 affective terms - unbalanced sparse distribution with an emphasis on sustained primarily *dark* aspects - the *last.fm* tag cloud includes: "mellow, love, chillout, dreamy, relax, romantic, sad, sexy, sleepy, smooth, soft, sweet"

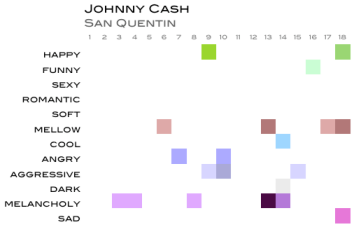


Fig. 5. LSA cosine similarity between the lyrics of "San Quentin" and 12 affective terms - unbalanced sparse distribution with an emphasis on sustained primarily *melancholy* aspects - the *last.fm* tag cloud includes: "cynical, prison, reflective, vice, visceral"

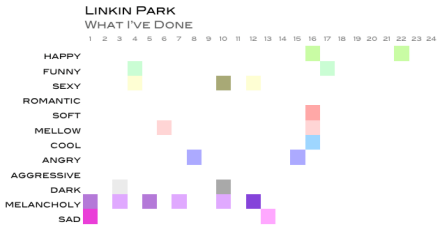


Fig. 6. LSA cosine similarity between the lyrics of "What i've done" and 12 affective terms - unbalanced sparse distribution biased towards *melancholy* aspects - the *last.fm* tag cloud includes: "energetic, love, memories, nice"

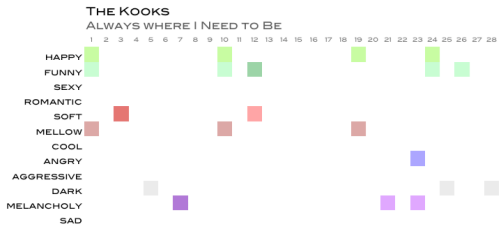


Fig. 7. LSA cosine similarity between the lyrics of "Always where i need to be" and 12 affective terms - unbalanced sparse distribution with emphasis on *funny, happy, mellow, soft* aspects - the *last.fm* tag cloud includes: "im in love with this song, makes me happy, cool, energetic, fun, high spirits, makes me wanna dance, relax, sounds like summer"

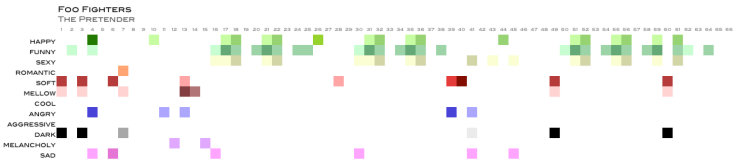


Fig. 8. LSA cosine similarity between the lyrics of "The pretender" and 12 affective terms - unbalanced sparse distribution with clusters of *funny, happy, sexy* against scattered *dark* and*soft,mellow* aspects - the *last.fm* tag cloud includes: "aggressive, angry, chill, cool, fun, kick ass, love, rebellious, soundtrack to your life"

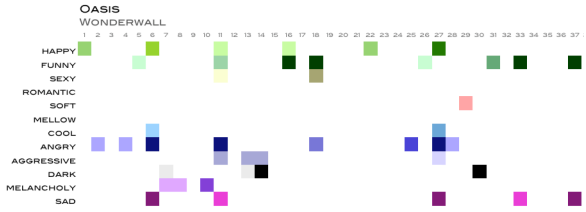


Fig. 9. LSA cosine similarity between the lyrics of 'Wonderwall" and 12 affective terms - unbalanced sparse distribution focused around *angry, sad, dark* complemented by *funny* and*happy* aspects - the *last.fm* tag cloud includes: "calm, chill, cool, emotional, love, mellow, nostalgia, romantic, sad"

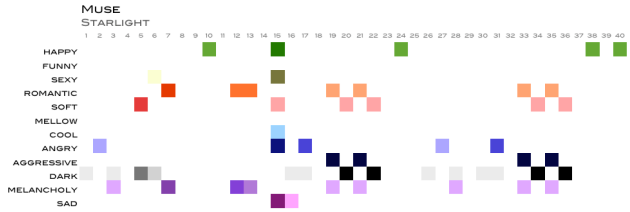


Fig. 10. LSA cosine similarity between the lyrics of "Starlight" and 12 affective terms - unbalanced sparse distribution concentrated around *dark, melancholy, angry, aggressive* aspects against *happy, romantic* elements - the *last.fm* tag cloud includes: "chill, emo, feelgood, hope, love, makes me happy, mellow, sexy, uplifting"

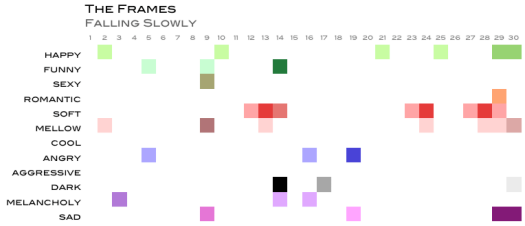


Fig. 11. LSA cosine similarity between the lyrics of "Falling Slowly" and 12 affective terms - centered distribution stressing sustained *soft* and *mellow* aspects - the *last.fm* tag cloud includes: "mellow, slow, emo, love, sad, sweet"

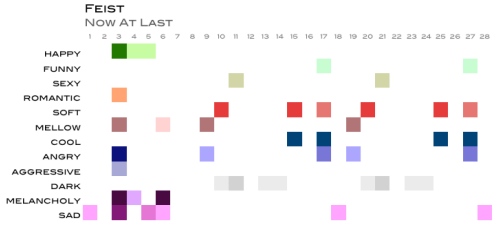


Fig. 12. LSA cosine similarity between the lyrics of "Now at last" and 12 affective terms - centered distribution stressing *soft* and *mellow* periodical aspects clustered with additional *cool* and *angry* elements - the *last.fm* tag cloud includes: "mellow, sad, chillout, melancholy, love, quiet, dreamy, relax, slow, soft, sweet, wistful"

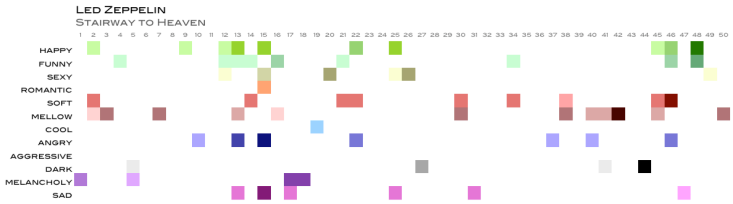


Fig. 13. LSA cosine similarity between the lyrics of "Stairway to heaven" and 12 affective terms - centered distribution emphasizing sustained *mellow* and *soft* aspects combined with the contrasts of *happy, funny* and *sad, melancholy* elements - the *last.fm* tag cloud includes: "chill, cool, melancholic"

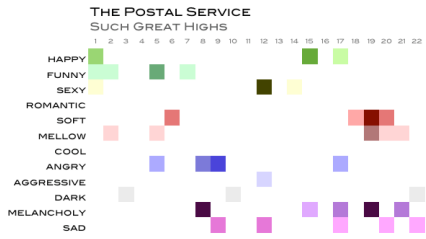


Fig. 14. LSA cosine similarity between the lyrics of "Such great heights" and 12 affective terms - uniform distribution offsetting bottom-heavy *melancholy, sad* against *soft, mellow* aspects, coupled with contribution of *sexy, funny, happy* elements - the *last.fm* tag cloud includes: "love, chillout, cool, emo, fun, happy, mellow, nice, sweet, upbeat"

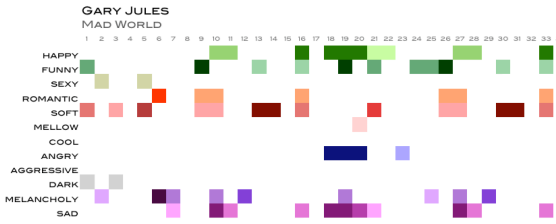


Fig. 15. LSA cosine similarity between the lyrics of "Mad world" and 12 affective terms - uniform distribution juxtaposing *happy, funny* against *sad* aspects, coupled with an equally strong contribution of *soft* elements - the *last.fm* tag cloud includes: "sad, melancholy, mellow, chillout, calm, dark, depressing, emotional, love, relax, slow, soft, touching"

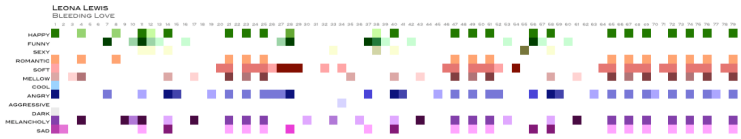


Fig. 16. LSA cosine similarity between the lyrics of "Bleeding love" and 12 affective terms - uniform distribution juxtaposing *happy, funny* against *angry, melancholy, sad* aspects, coupled with clusters of *soft, mellow, romantic* elements - the *last.fm* tag cloud includes: "love, emotional, lifelike, melancholy, mellow, romantic, sad, sexy, smooth, sweet"

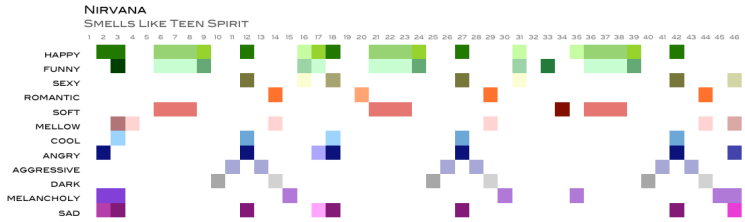


Fig. 17. LSA cosine similarity between the lyrics of "Smells like teen spirit" and 12 affective terms - uniform distribution contrasting *happy, funny* with *sad* and *angry* aspects - the *last.fm* tag cloud includes: "cool, love, melancholic, nostalgic, "

4 Conclusion

While we have here only analyzed twenty songs our first results indicate that it is possible to retrieve a fraction of the emotional elements contributing to user descriptions of songs, by applying LSA to extract latent semantics from the lyrics using a selection of frequently occurring *last.fm* affective tags. When comparing the LSA accumulated emotional components extracted from the lyrics with the actual user defined tag clouds of the corresponding songs at *last.fm* they appear to a large extent overlapping or complementary related to the affective terms. Analyzing the emotional components reflected in the lyrics over the entire duration of a song, we seem able to separate these aspects in plots defined by the sparsity and character of the distribution.

Considering the amount of structure in the song lyrics which emerge from the above matrix examples, coupled with the neuroimaging results indicating that music and language are processed in interaction, we speculate that these

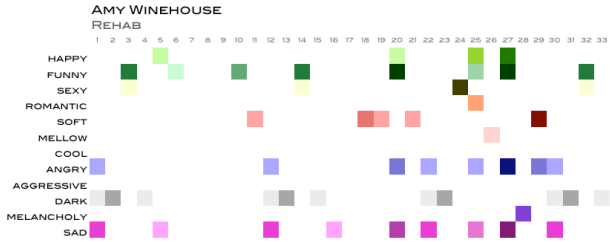


Fig. 18. LSA cosine similarity between the lyrics of "Rehab" and 12 affective terms - uniform distribution approaching a more sparse concentration towards the edges juxtaposing *funny, happy* against *dark, sad* as well as complementary *angry* aspects - the *last.fm* tag cloud includes: "funk, alcohol, chillout, cool, fun, happy, love, mellow, sexy, smooth"

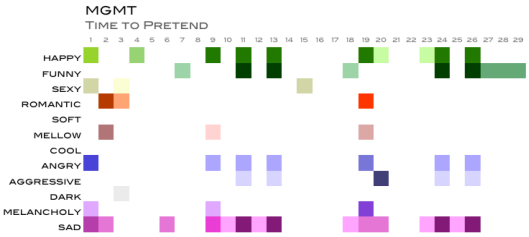


Fig. 19. LSA cosine similarity between the lyrics of "Time to pretend" and 12 affective terms - balanced sparse distribution offsetting clusters of *happy, funny* against sustained *sad* aspects - the *last.fm* tag cloud includes: "catchy, drugs, happy, infectious, witty, addictive, bittersweet, cool, dreamy, love, nostalgic, pensive"

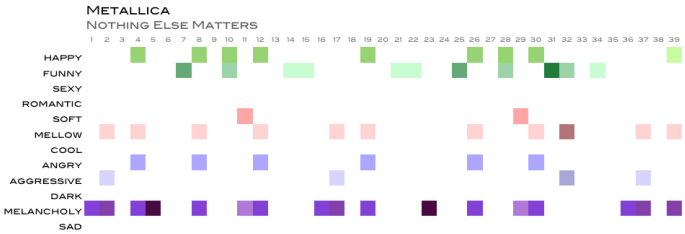


Fig. 20. LSA cosine similarity between the lyrics of "Nothing else matters" and 12 affective terms - balanced sparse distribution of *happy, funny* juxtaposed against sustained *melancholy* as well as *mellow* aspects - the *last.fm* tag cloud includes: "chillout, dark, melancholic, relax, sad"

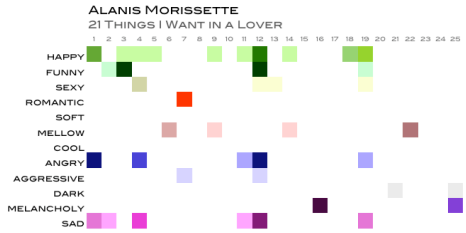


Fig. 21. LSA cosine similarity between the lyrics of "21 Things i want in a lover" and 12 affective terms - balanced sparse distribution of *happy, funny* complemented with *sad* and *angry* aspects - the *last.fm* tag cloud includes: "attitude, emotions, if love had a soundtrack, independent and in-your-face, intelligent, kickass, makes me laugh, probing, witty"

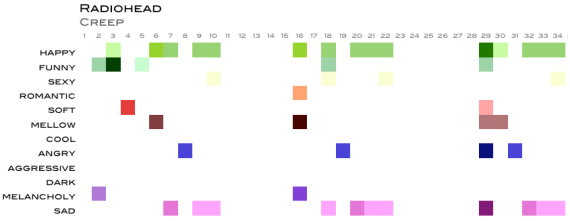


Fig. 22. LSA cosine similarity between the lyrics of "Creep" and 12 affective terms - balanced sparse distribution emphasizing the extremes of *happy, funny* complemented with *sad* and *angry* aspects

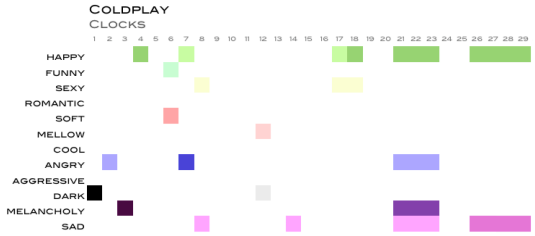


Fig. 23. LSA cosine similarity between the lyrics of "Clocks" and 12 affective terms - balanced sparse distribution juxtaposing the extremes of *happy* with *sad* and *melancholy* combined with *angry* aspects - the *last.fm* tag cloud includes: "chill, chillout, cool, dreamy, love, melancholic, mellow, nostalgic, relax, soft"

emotional components might provide a basis for modeling patterns related to how emotions arise based on the underlying structures forming our expectations. We suggest that a fraction of these components might be retrieved as latent semantics by using affective terms as sensors in a semantic space, and hypothesize that LSA might be applied to extract structural patterns from song lyrics as a basis for automatically generating emotional playlists. The LSA matrix examples of correlation between song lyrics and affective terms seem to indicate that even if we turn off the sound both the emotional context of the texts as well as overall formal structural elements in the music can to a certain extent be extracted from the latent semantics.

References

1. L. B. Meyer: *Meaning in music and information theory*, Journal of Aesthetics and Art Criticism, Vol.15, pp. 412-424, 1957
2. D. Temperlie: *Music and probability*, MIT Press, 2007
3. D. Huron: *Sweet anticipation: Music and the psychology of expectation*, MIT Press, 2006
4. R. Jackendoff and F. Lerdahl, *The capacity for music: what is it, and what's special about it?*, Cognition, pp. 33-72, 2006
5. C.L. Krumhansl, *Music: A link between cognition and emotion*, Current Directions in Psychological Science pp.35-55, 2002
6. D.J. Leviten and V. Menod: *Musical structure is processed in "language" areas of the brain: a possible role for Brodmann Area 47 in temporal coherence*, NeuroImage, pp. 2142-2152, 2003
7. D Schön, R.L. Gordon, and M. Besson: *Musical and linguistic processing in song perception*, Annals of the New York Academy of Sciences, pp. 71-81, 2005
8. I. Peretz, R. Gagnon, and S. Hbert: *Singing in the brain: Insights from cognitive neuropsychology*, Music Perception, pp. 373-390, 2004
9. I. Peretz, M. Radeau, and M. Arguin: *Two-way interactions between music and language: Evidence from priming recognition of tune and lyrics in familiar songs*, Memory & Cognition, pp. 42-52, 2004
10. M. Levy and M. Sandler, *A semantic space for music derived from social tags*, Proceedings of the 8th International Conference on Music Information Retrieval, pp. 411-416, 2007
11. X. Hu, M. Bay and S.J. Downie: *Creating a simplied music mood classication ground-truth set*, Proceedings of the 8th International Conference on Music Information Retrieval, pp. 309-310, 2007
12. L.K. Landauer and S.T. Dumais: *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge*, Psychological Review, pp. 211-240, 1997
13. C.E. Osgood, G.J. Suci and H.P. Tannenbaum: *The measurement of Meaning*, University of Illinois Press, 1957

Appendix D

Extracting Moods from Songs and BBC Programs Based on Emotional Context.

Published in the *"International Journal of Multimedia Broadcasting", Volume 2008, Article ID 289837.*

Hindawi Publishing Corporation
International Journal of Digital Multimedia Broadcasting
Volume 2008, Article ID 289837, 12 pages
doi:10.1155/2008/289837

Research Article

Extracting Moods from Songs and BBC Programs Based on Emotional Context

Michael Kai Petersen and Andrius Butkus

Department of Informatics and Mathematical Modeling, Technical University of Denmark, Richard Petersens Plads, Building 321, 2800 Kongens Lyngby, Denmark

Correspondence should be addressed to Michael Kai Petersen, mkp@imm.dtu.dk

Received 2 March 2008; Revised 2 July 2008; Accepted 4 August 2008

Recommended by Harald Kosch

The increasing amounts of media becoming available in converged digital broadcast and mobile broadband networks will require intelligent interfaces capable of personalizing the selection of content. Aiming to capture the mood in the content, we construct a semantic space based on tags, frequently used to describe emotions associated with music in the *last.fm* social network. Implementing latent semantic analysis (LSA), we model the affective context of songs based on their lyrics, and apply a similar approach to extract moods from BBC synopsis descriptions of TV episodes using TV-Anytime atmosphere terms. Based on our early results, we propose that LSA could be implemented as machinelearning method to extract emotional context and model affective user preferences.

1. INTRODUCTION

When both digital broadcast streams and the content itself are adapted to the small screen size of handheld devices, it will literally translate into hundreds of channels featuring rapidly changing mobisodes and location-aware media, where it might no longer be feasible to select programs by scrolling through an electronic program guide. In order to automatically filter media according to personalized preferences, this will require metadata which not only defines traditional genre categories but also incorporates parameters capturing the changing mobile usage contexts. Since 2005, the broadcaster BBC has made their program listings available as XML formatted TVA TV-Anytime [1] metadata, which allows for describing media using complementary aspects, such as content genre, format, intended audience, intention, or atmosphere. We have previously in a related paper [2] analyzed how especially atmosphere metadata describing emotions may facilitate identifying programs that might be perceived as similar even though they belong to different genre categories. Also in music it appears that despite the often idiosyncratic character of tags, defined by hundred thousands of users in social networks like *last.fm*, people tend to agree on the affective terms they attach to describe music [3, 4]. A mounting question might therefore be: could we possibly apply machine learning techniques to extract emotional aspects associated with media in order to model our perception, and thus facilitate an affective categorization which goes beyond traditional divides of genres?

2. RELATED WORKS

In usage scenarios involving DVB-H mobile TV, where shifting between a few channels might be even more time-consuming than watching the actual mobisode, new text mining approaches to content-based filtering have been suggested as a solution. Reflecting preferences for categories like "fun," "action," "thrill," or "erotic," topics and emotions are extracted from texts describing the programs and incorporated into the EPG electronic program guide data as a basis for generating user preferences [5]. In broadcast context, a similar approach has been implemented to extract both textual and visual concepts for automatic categorization of TV ad videos based on probabilistic latent semantic analysis (pLSA) [6]. As a machine learning method similar

to latent semantic analysis (LSA) [7], it captures statistical dependencies among distributions of visual objects or brand names, and thus enables unsupervised categorization of semantic concepts within the content. Recent neuroimaging experiments, focused on visualizing human brain activity reflecting the meaning of nouns, have demonstrated a direct relationship between the observed patterns in brain scans of regions being activated, and the statistics of word cooccurrence in large collections of documents. The distinct patterns of functional magnetic resonance images (fMRIs) triggered by specific terms seem not only to cause similar brain activities across different individuals [8], but also makes it possible to predict which voxels in the brain will be activated according to semantic categories based on word cooccurrence in a large text corpus [9]. Or in other words, the way LSA simulates text comprehension by modelling the meaning of words as the sum of contexts in which they occur appears to have neural correlates.

Over the past decade, advances in neuroimaging technologies enabling studies of brain activity have established that musical structure to a larger extent than previously thought is being processed in "language" areas of the brain [10]. Neural resources between music and language appear to be shared both in syntactic sequencing and also semantic processing of patterns reflecting tension and resolution [11–13], adding support for findings of linguistic and melodic components of songs being processed in interaction [14]. Similarly, there appears to be an overlap between language regions in the brain and mirror neurons, which transfer sensory information of what we perceive by reenacting them on a motor level. The mirror neuron populations mediate the inputs across audiovisual modalities and the resulting sensory-motor integrations are represented in a similar form, whether they originate from actions we observe in others, only imagine or actually enact ourselves [15, 16]. This has led to the suggestion that our empathetic comprehension of underlying intentions behind actions, or the emotional states reflected in sentences and melodic phrases are based on an imitative reenactment of the perceived motion [17].

Aspects of musical affect have been the focus of a wide field of research, ranging from how emotions arise based on the underlying harmonic and rhythmical hierarchical structures forming our expectations [18–20], to how we consciously experience these patterns empathetically as contours of tensions and release [21], in turn triggering physiological changes in heart rate or blood pressure as has been documented in numerous cognitive studies of the links between music and emotions [22]. But when listening to songs our emotions are not only evoked by low-level cognitive representations but also exposed to higher level features reflecting the words which make up the lyrics. Studies on retrieving songs from memory indicate that lyrics and melody appear to be recalled from two separate versions: one storing the melody and another containing only the text [23], while further priming experiments indicate that song memory is not organized in strict temporal order, but rather that text and tune intertwine based on reciprocal connections of higher-order structures [24].

Taking the above findings into consideration, could we possibly extract affective components from textual representations of media like song lyrics, and model them as patterns reflecting how we emotionally perceive media? Applying LSA as a machine learning method to extract moods in both song lyrics and synopsis descriptions of BBC programs, we describe in the following sections, the methodology used for extracting high level representations of media using emotional tags, the early results retrieved when mapping emotional components of song lyrics and synopsis descriptions, and conclude with a discussion of the potential for automatically generating affective user preferences as a basis for mood-based recommendation.

3. EMOTIONAL TAG SPACE

When investigating how unstructured metadata can be used to describe media, the social music network *last.fm* provides an interesting case. The affective terms which are frequently chosen as tags by *last.fm* users to describe the emotional context of songs seem to form clusters around primary moods like mellow, sad, or more agitated feelings like angry and happy. This correlation between social network tags and the specific music tracks they are associated with has been used in the music information retrieval community to define a simplified mood ground-truth, reflecting not just the words people frequently use when describing the perceived emotional context, but also which tracks they agree on attaching these tags to [3, 4]. We have selected twelve of these frequently used tags for creating an emotional semantic space. Drawing on standard psychological parameters for emotional assessment, we map these affective terms along the two primary dimensions of *valence* and *arousal* [25], and use these two axes to outline an emotional plane for dividing them within an affective semantic space containing four groups of frequently used *last.fm* tags:

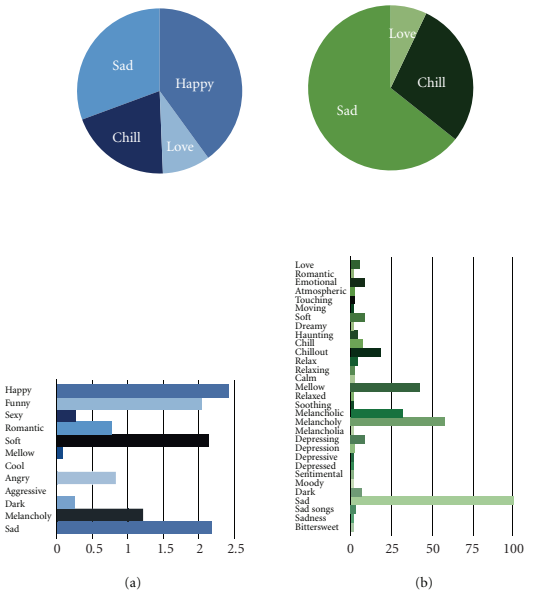
(i) *happy, funny, sexy;*

(ii) *romantic, soft, mellow, cool;*

(iii) *angry, aggressive;*

(iv) *dark, melancholy, sad.*

Within this emotional plane, the dimension of *valence* describes how pleasant something is along an axis going from positive to negative associated with words like happy or sad, whereas *arousal* captures the amount of involvement ranging from passive states like mellow and sad to active aspects of excitation as reflected in tags like angry or happy. Applying the selected *last.fm* tags as emotional buoys to define a semantic plane of psychological valence and arousal dimensions, we apply latent semantic analysis (LSA) to assess the correlation between the lyrics and each of the selected affective terms. Applying these affective terms as markers also enables us to compare the LSA-retrieved values against the actual tags users have applied in the *last.fm* tag clouds associated with the songs in our analysis. Additionally, when analyzing the synopsis descriptions of BBC programs we have complemented the *last.fm* tags with a large number of TV-Anytime atmosphere terms similarly used as emotional



(a) (b)

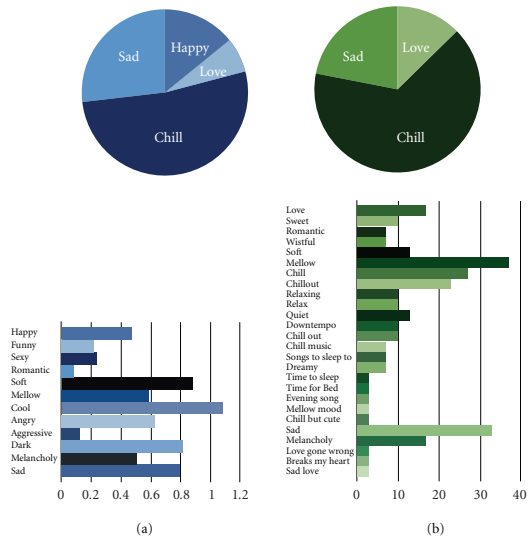
FIGURE 1: Accumulated LSA correlation between (a) the lyrics of the song "Nothing else matters" and 12 affective terms, compared to (b) the actual user-defined emotional tags at last.fm.

buoys. Though the two sets of markers are clearly affected differently by the synopsis, a comparison shows that despite the higher degree of detail in the TV-Anytime vocabulary, the overall emotional context is reflected similarly by the *last.fm* tags and the atmosphere terms. Or in other words, the *last.fm* and TV-Anytime markers provide different granularities for capturing emotions but the larger tendencies in the resulting patterns remain the same.

As a machine learning technique, LSA extracts meaning from paragraphs by modelling the usage patterns of words in multiple documents and represent the terms and their contexts as vectors in a high-dimensional space. The basis for assessing the correlations between lyrics and emotional words vectors in LSA is an underlying text corpus consisting of a large collection of documents which provides the statistical basis for determining the cooccurrence of words in multiple contexts. For this experiment, we chose the frequently implemented standard *TASA* text corpus, consisting of the 92409 words found in 37651 texts, novels, news articles, and other general knowledge reading material that American students are exposed to up to the level of their 1st year in college. The frequency at which terms appear and the phrases wherein they occur are defined in a matrix with rows made up of words and columns of documents. Many of the cells made up by rows and columns contain only zeroes, so in order to retain only the most essential features, the dimensionality of the original sparse matrix is reduced to around 300 dimensions. This makes it possible to model the semantic relatedness of song lyrics and affective terms as vectors, with values toward 1 signifying degrees of similarity between the items and low or minus values typically around 0.02 signifying a random lack of correlation. In this semantic space lines of lyrics or emotional words which express the same meaning will be represented as vectors that are closely aligned, even if they do not literally share any terms. Instead, these terms may cooccur in other documents describing the same topic, and when reducing the dimensionality of the original matrix, the relative strength of these associations can be represented as the cosine of the angle between the vectors.

4. RESULTS: SONG LYRICS

Whereas the user-defined tags at *last.fm* describe a song as a whole, we aim to model the shifting contours of tension and release which evoke emotions, and therefore project each of the individual lines of the lyrics into the semantic space. Analyzing individual lines on a timescale of seconds also reflects the cognitive temporal constraints applied by our brains in general when we bind successive events into perceptual units [26]. We perceive words as



(a) (b)

Figure 2: Accumulated LSA correlation between (a) the lyrics of the song "Now at last" and 12 affective terms, compared to (b) the actual user-defined emotional tags at last.fm.

successive phonemes and vowels on a scale of roughly 30 milliseconds, which are in turn integrated into larger segments with a length of approximately 3 seconds. We thus assume that lines of lyrics consisting of a few words each correspond to one of these high-level perceptual units. Viewed from a neural network perspective, projecting the lyrics into a semantic LSA space line by line, could also in a cognitive sense be interpreted as similar to how mental concepts are constrained by the amount of activation among the neural nodes representing events and associations in our working memory [27]. In that respect, the cooccurrence matrix formed by the word frequencies of *last.fm* tags and song lyrics might be understood as corresponding to the strengths of links connecting nodes in a mental model of semantic and episodic memory.

4.1. Accumulated emotional components

Projecting the lyrics of thirty songs selected from the weekly top track charts at *last.fm*, we compute the correlation between lyrics and tags against each of the twelve affective terms used as markers in the LSA space, while discarding cosine values below a threshold of 0.09. And in order to compare the retrieved LSA correlation values of lyrics and affective terms against the user-defined tags attached to the song at *last.fm*, we sum up the accumulated LSA values retrieved from each line of the lyrics.

Taking the song "Nothing else matters" as an example, the user defined tags attached to the song as at *last.fm*, include less frequently used tags like *love, love songs, chill, chillout, relaxing, relax, memories,* and *melancholic* which are not among the markers we used for our LSA analysis. We therefore subsequently combine these tags into larger segments of tags in order to facilitate a direct comparison with the LSA-retrieved values (Figure 1). Comparing the accumulated LSA values of emotional components against the user-defined tags at *last.fm*, the terms *melancholy,* and *melancholic*, which describe the most dominant emotions in the tag cloud, could be understood as captured by the affective term *sad* in the LSA analysis. Similarly, if interpreting *love* from the *last.fm* tag cloud as associated with the term *happy* (based on a cosine correlation of 0.56 between the words love and happy), the LSA analysis could be understood to retrieve also aspects of this emotion. Likewise, if *chill* in the *last.fm* tag cloud is understood as associated with *soft* and *mellow* (based on cosine correlations of 0.36 and 0.35, resp.), the LSA analysis also here appears to capture that mood.

Applying a similar approach to a set of thirty songs, we grouped semantically close *last.fm* tags into larger segments consisting of *sad, happy, love,* and *chill* aspects to facilitate a comparison with the LSA-derived correlations between song lyrics and the selected affective terms. Though there is an overlap between the retrieved LSA values and user-defined

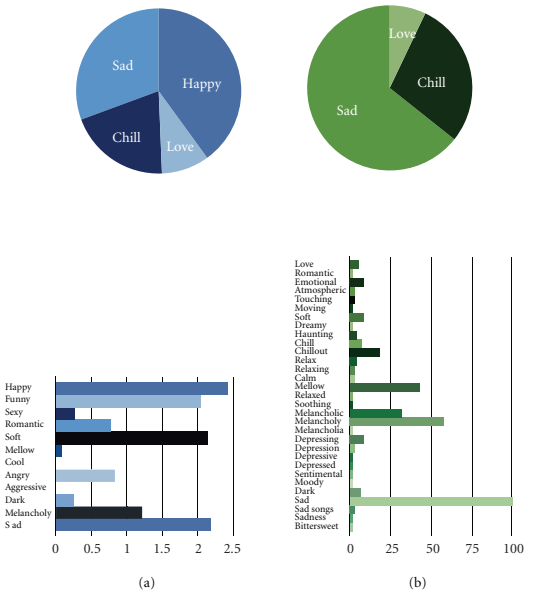


FIGURE 3: Accumulated LSA correlation between (a) the lyrics of the song "Mad world" and 12 affective terms, compared to (b) the actual user-defined emotional tags at last.fm.

last.fm tags in most of the songs, there is no overall significant correlation between LSA-retrieved values and the exact distribution of tags in the user-defined *last.fm* tag clouds. Essentially, the individual tags in a cloud are "one size fits all" and apply to the song as a whole, whereas the LSA correlation between lyrics and semantic markers reflects the changing degrees of affinity between the song lines and affective components over time. But for a third of the set of songs, as exemplified by "Now at last" (Figure 2), the distribution of *last.fm* tags resembled the LSA values if grouped into larger segments. While in the remaining two thirds of the set of songs, as exemplified by the song "Mad World" (Figure 3), the overall distribution in *last.fm* tags while clearly overlapping remain overly biased toward *sad* type of components.

4.2. Distribution of emotional components

Instead of grouping the emotional components into larger segments, we subsequently maintained the LSA values retrieved from each of the individual lines in the lyrics, and proceeded by plotting the values over time to provide a view of the distribution of emotional components. The plots can be interpreted as mirroring the structure of patterns of changing emotions in the songs along the horizontal axis. Vertically, the color groupings indicate which of the aspects of valence and arousal are triggered by the lyrics as well as their general distribution in relation to each other. Any color will signify an activation beyond the cosine similarity threshold level of 0.09, and the amount of saturation from light to dark signifies the degree of correlation between the song lyrics and each of the affective terms. The contribution of each emotional component apparent in the overall LSA values of the lyrics can be made out when considering their distribution as single pixels over time triggered by the individual lines in each of the songs. When analyzing which emotional components appear predominant and overall contribute the most, the LSA plots can roughly be grouped into three categories which can be characterized as *unbalanced distributions*, *centered distributions,* and *uniform distributions*.

Going back to the song "Nothing else matters," Figure 4, the plot exemplifies the first *unbalanced* category by in this case having a bottom-heavy distribution of emotional components biased toward *melancholy*. The below curve of accumulated LSA values indicates the contribution of each component over the entire song, where the significant aspects of *melancholy* are clearly separated from the other components.

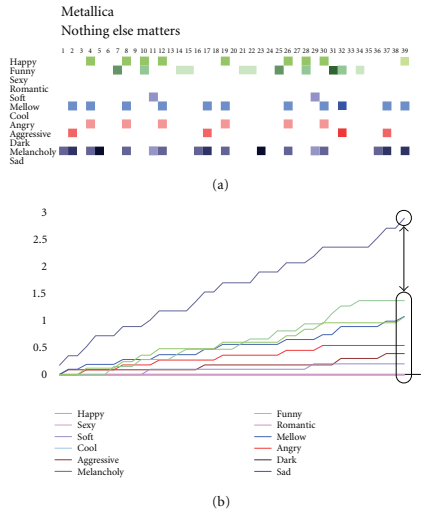


Figure 4: LSA correlation between (a) the lyrics of the song "Nothing else matters" and 12 affective terms, with (b) accumulated values plotted over the entire length of the song.

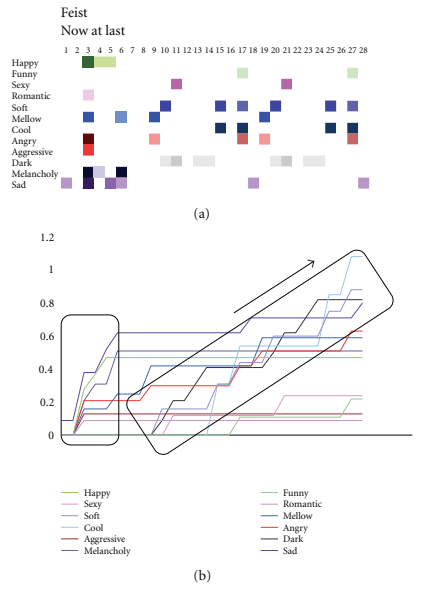


Figure 5: Summed up values of LSA correlation between (a) the lyrics of the song "Now at last" and 12 affective terms, with (b) accumulated values plotted over the entire length of the song.

The *centered distribution* distribution as found in "Now at last" (Figure 5) shows a lack of the more explicit emotions like "happy" or "sad" apart from the very beginning, while instead the main contribution throughout the song comes from more passive "mellow" and "soft" aspects. In contrast to the former example, the below curves of accumulated emotional contributions reflect a pattern combining the activation of "happy" or "sad" elements which remain at the initial level, whereas the more passive aspects "mellow" and "soft" are continuously accumulating throughout the song.

A *uniform distribution* of a wide range of simultaneous emotional components is exemplified by "mad world," Figure 6, simultaneously juxtaposing emotional areas around "happy" against "sad" components. This pattern can also be made out in the below curves, where additionally the sudden steep increase in accumulated values starting roughly a third into the song also illustrates how the emotional components reflect the overall structure in the song.

The overall saturation defining the amount of correlation between lyrics and emotional markers, as well as the distributional patterns of emotional components throughout the songs seem consistent. Lyrics that appear more or less saturated in relation to the emotional markers used for the LSA analysis remain so over the entire song. The distributional patterns of emotional elements seem throughout the songs to form consistent schemas of contrasting elements, which appear to form sustained lines or clusters that are preserved as pattern once initiated. We suggest that these elements form bags of features, which could be used to categorize and infer patterns as a basis for building emotional playlists. From these features, general patterns emerge, as in the distributions of emotional components in the songs "Wonderwall" and "My Immortal," Figure 7, which appear similar due to a sparsity of central aspects like "soft," while instead emphasizing the outer edges by juxtaposing elements around "happy" against "sad." The opposite character can be seen in the distributions of central elements stressed in the songs "Falling slowly" and "Stairway to heaven," Figure 8, which underline the aspects of "soft" and "mellow" at the expense of "happy" and "sad." Whereas these elements in the songs "Everybody hurts" and "Smells like teen spirit," Figure 9, appear as structural components grouped into clusters, either providing a strong continuous activation of complementary feelings or juxtaposing these emotional components against each other.

5. RESULTS: BBC SYNOPSIS

Repeating the approach, but this time to extract emotions from texts describing TV programs, we take a selection of short BBC synopses as input, and compute the cosine

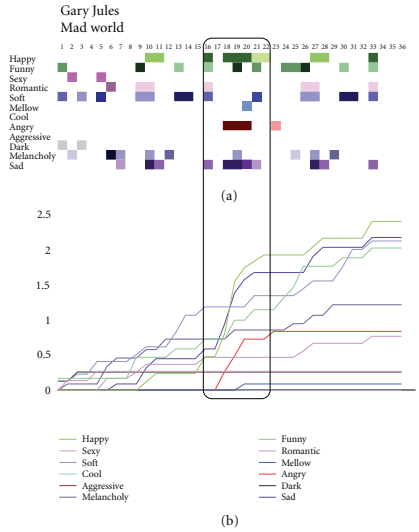


Figure 6: Summed up values of LSA correlation between (a) the lyrics of the song "Mad world" and 12 affective terms, with (b) accumulated values plotted over the entire length of the song.

similarities between a synopsis text vector and each of the selected *last.fm* emotional words. While the previously analyzed lyrics could be seen as integral parts of the original media, a synopsis description is clearly not. It only provides a brief summary of the program, but it nevertheless offers an actual description complementary to the associated *TV-Anytime* metadata genres. We initially analyzed a number of standalone synopsis descriptions to see if would be possible to capture emotional aspects of the BBC programs.

An analysis of the program "News night," based on the short description: *News in depth investigation and analysis of the stories behind the day('s) headline*, triggers the tags "funny" and "sexy" which might not immediately seem a fitting description, probably caused by these emotional terms being directly correlated with the occurrence of the words stories and news within the synopsis. The atmosphere of the lifestyle program "Ready Steady Cook!" might be somewhat better reflected in the synopsis: *Peter Davidson and Bill Ward challenge celebrity chefs to create mouth watering meals in minutes*, which triggers the tag "romantic" as associated with meals. Another singular emotion can be retrieved from the documentary "I am a boy anorexic," which based on the synopsis: *Documentary following three youngsters struggling to overcome their obsessive relationship with food as they recover inside a London clinic and then return to the outside world*, triggers the affective term "dark." We find a broader emotional spectrum reflected in the lifestyle program "The flying gardener" described by the text: *The flying gardener Chris travels around by helicopter on a mission to find Britain('s) most inspirational gardens. He helps a Devon couple create a beautiful spring woodland garden. Chris visits impressive local gardens for ideas and reveals breathtaking views of Cornwall from the air*. The synopsis triggers a concentration of passive pleasant *valence* elements related to the words "soft, mellow" combined with "happy." In this context also the tag "cool" comes out as it has a strong association to the word air contained in the synopsis, while the activation of the tag "aggressive" appears less explainable. This cluster of pleasant elements is lacking in the LSA analysis of the program "Super Vets" which instead evokes a strong emotional contrast based on the text: *At the Royal Vet College Louis the dog needs emergency surgery after a life threatening bleed in his chest and the vets need to find out what is causing the cat fits*, where both pleasant and unpleasant active terms like "happy" and "sad" stand out in combination with strong emotions reflected by the tag "romantic." And as can be seen from programs like "The flying gardener" and "Super Vets" (Figure 10), the correlation between the synopsis and the chosen tags might often trigger both complementary elements as well as contrasting emotional components.

We proceeded to explore whether we could sum up a distinct pattern reflecting an emotional profile pertaining to a TV series, by accumulating the LSA values of correlation between synopsis texts and emotional tags over several episodes. Similar to our previous approach when analyzing lyrics, where we held the LSA results against the user defined *last.fm* tag clouds, we here compare the LSA values of the synopsis against the *TV-Anytime* atmosphere genres used in the BBC metadata. This classification scheme offers 53 different terms which might be included in the genre metadata to express the atmosphere or perceived emotional response when watching a program. Projecting the synopsis descriptions against 53 *TV-Anytime* terms, used as emotional markers in the LSA analysis, allows for defining more differentiated patterns. At the same time also projecting the BBC synopsis against the previously used *last.fm* tags in the LSA analysis, makes it possible to compare to what extent the choice of using either *TV-Anytime* atmosphere terms or *last.fm* tags as emotional markers in the semantic space is influencing the results.

For analyzing the emotional context in a sequence of synopsis descriptions of the same program, we chose the soap "East Enders," the comedy "Two pints of lager," and sci-fi series "Doctor Who." Initially, plotting the LSA analysis of the soap "East Enders" and comedy "Two pints of lager" against 12 *last.fm* tags (Figures 1 and 2, increased color saturation corresponds to degree of correlation), the distributions of emotional components appear unbalanced in both cases. But whereas the soap has a bottom-heavy bias toward "sad" and "angry" outweighing "happy," the balance is reversed in the comedy which shifts towards predominantly "happy" and "funny" complemented by "soft" and "mellow" aspects. Overall, the distribution in "East Enders" is much more dense and emotionally saturated as

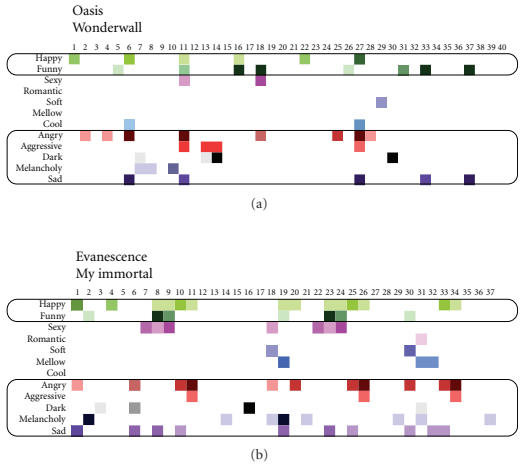


FIGURE 7: Pairwise comparison of patterns reflecting LSA correlation values in the lyrics of the songs (a) "Wonderwall", and (b) "My immortal" against 12 affective terms.

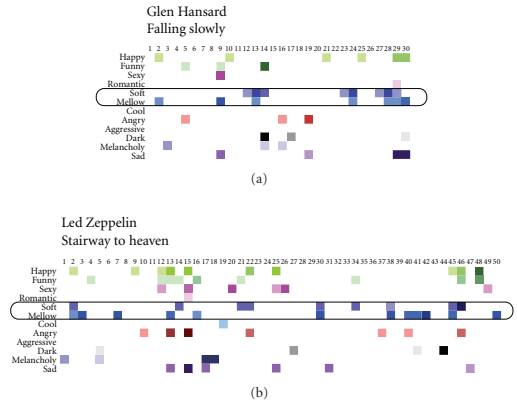


FIGURE 8: Pairwise comparison of patterns reflecting LSA correlation values in the lyrics of the songs (a) "Falling slowly", and (b) "Stairway to heaven" against 12 affective terms.

exemplified in elements like "angry" reflecting high arousal. In contrast, the lighter character of "Two pints of lager" comes out in the clustering of positive valence elements such as "happy" and "funny," coupled with a general sparsity of excitation within the matrix.

As a second step, projecting the synopsis descriptions against the 53 *TV-Anytime* atmosphere terms of course results in more differentiated patterns. Users at *last.fm* frequently describe tracks as "angry" but as music is rarely described as scary, feelings of fear are lacking. Otherwise,

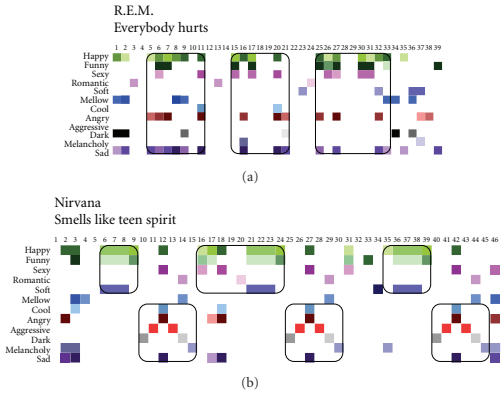


FIGURE 9: Pairwise comparison of patterns reflecting LSA correlation values in the lyrics of the songs (a) "Everybody hurts", and (b) "Smells like teen spirit" against 12 affective terms.

so with the *TV-Anytime* metadata which also captures these aspects in a synopsis with atmosphere terms like "terrifying." Some of these elements are essential for describing the content as is evident in the sci-fi series "Doctor Who," Figure 13. Lacking words for these feelings, the *last.fm* tags "Melancholy" and "dark" are triggered, whereas it takes the increased resolution of the *TV-Anytime* atmosphere terms to capture the equally "spooky" and "silly" aspects.

Altogether *TV-Anytime* adds a large number of terms, which rather than describing emotions capture attitudes or perceived responses like "stylish" or "compelling," and as such trigger vast amounts of elements contributing to the atmosphere. In "East Enders" adding elements like "frantic" and "exciting" to the pattern. Similarly, the larger number of comical elements exemplified by words like "crazy, silly," or "wacky" provides a much higher emotional granularity in the description of "Two pints of lager". However, the overall bias toward positive or negative valence and arousal within the distributions seem largely preserved, independent of whether *last.fm* or *TV-Anytime* terms are used as emotional markers in the LSA analysis.

Comparing the emotional components retrieved from the LSA analysis of the synopsis texts against the actual *TV-Anytime* atmosphere terms in the BBC metadata, they seem to be largely in agreement. The comedy has been indexed as "humorous, silly, irreverent, fun, wacky, crazy," while based on the synopsis texts alone, most of these components also come out in the LSA analysis. In the case of the soap "East Enders," the episodes are annotated as "gripping, gritty, gutsy." Although these terms are also triggered from the synopsis texts, these aspects might be even more reflected in the stark accumulated contrasts of "happy" and "sad" components retrieved by the LSA analysis. Similarly, in "Doctor Who" the actual *TV-Anytime* atmosphere terms applied in the BBC metadata *spooky, exciting* are also captured, while the grey patterns of perceived responses seem to add a lot more nuances to this description.

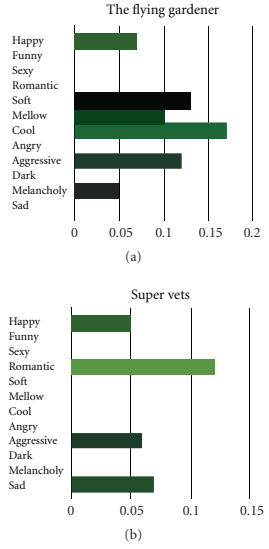


FIGURE 10: LSA cosine similarity between the synopsis descriptions of "The flying gardener" and "Super Vets" against 12 frequently used *last.fm* affective terms.

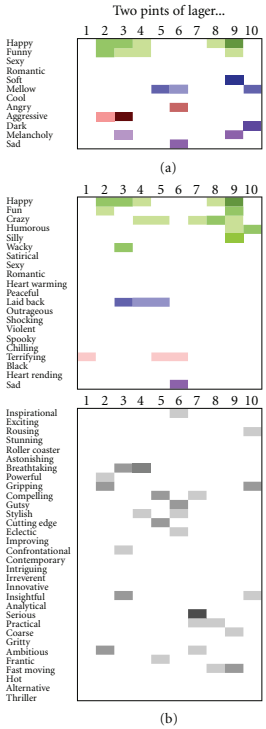


Figure 11: LSA correlation values of 10 episodes of (a) "Two Pints of lager" against 12 last.fm tags, and (b) 53 tva atmosphere terms.

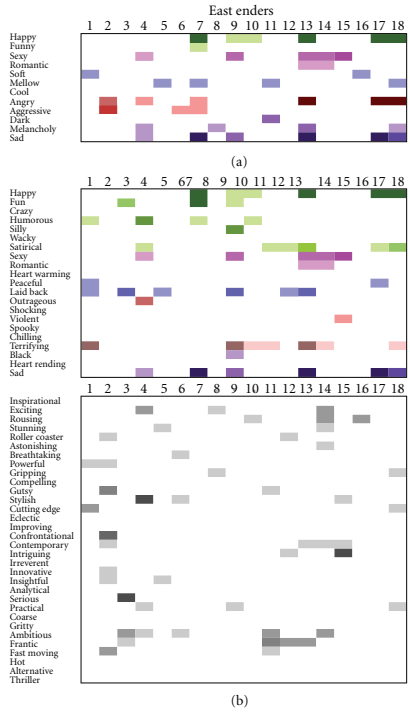


Figure 12: LSA correlation values of 18 episodes of (a) "East Enders" against 12 last.fm tags, and (b) 53 tva atmosphere terms.

6. CONCLUSIONS

Projecting BBC synopsis descriptions into an LSA space, using both *last.fm* tags and *TV-Anytime* atmosphere terms as emotional buoys Figures 11–13, we have demonstrated an ability to extract patterns reflecting combinations of emotional components. While each synopsis triggers an individual emotional response related to a specific episode, general patterns still emerge when accumulating the LSA correlation between synopsis and emotional tags over consecutive episodes, which enables us to differentiate between a comedy and a soap based on textual descriptions alone. Applying more semantic markers in the analysis allows for capturing additional elements of atmosphere in terms of perceived attitudes or responses to the media being consumed. However, the overall balance of affective components reflecting the media content seems largely preserved, independent of whether *last.fm* or *TV-Anytime* terms are used as emotional markers in the LSA analysis.

Moving beyond the static LSA analysis of consecutive synopsis descriptions, plotting the components over time might provide a basis for modelling the patterns of emotions evolving when we perceive media. We hypothesize that these emotional components reflect compositional structures perceived as patterns of tension and release, which form the dramatic undercurrents of an unfolding story line. As exemplified in the plots of song lyrics each matrix column corresponds to a time window of a few seconds, which is also the approximate length of the high-level units from which we mentally construct our perception of continuity within time [26]. Interpreted in that context, we suggest that the LSA analysis of textual components within a similar size of time window is able to capture a high level representation of the shifting emotions triggered by the media. Or from a cognitive perspective, the dimensionality reduction enforced by LSA might be interpreted as a simplified model of how mental concepts are constrained by the strengths of links connecting nodes in our working memory [27].

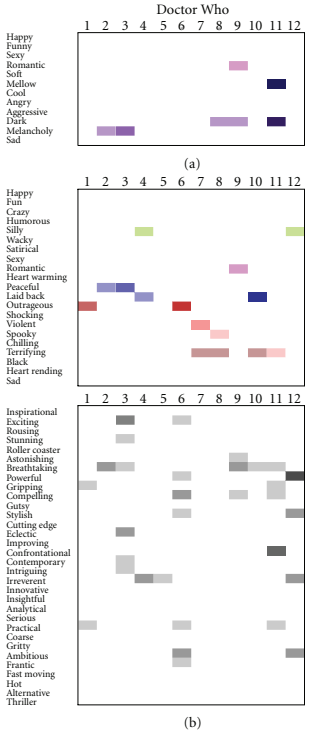


FIGURE 13: LSA correlation values of 12 episodes of (a) "Doctor Who" against last.fm tags, and (b) 53 tva atmosphere terms.

Finding that the emotional context of media can be retrieved by using affective terms as markers, we propose that LSA might be applied as a basis for automatically generating mood-based recommendations. It seems that even if we turn off both the sound and the visuals, emotional context as well as overall formal structural elements can still be extracted from media based on latent semantics.

REFERENCES

[1] ETSI, TV-Anytime. Part 3. Metadata 1. Sub-part 1. Part 1—Metadata schemas, TS 102822-3-1, 2006.

[2] A. Butkus and M. K. Petersen, "Semantic modelling using TV-anytime genre metadata," in *Proceedings of the 5th European Conference on Interactive TV: A Shared Experience (EuroITV '07)*, P. Cesar, K. Chorianopoulos, and J. F. Jensen, Eds., vol. 4471 of *Lecture Notes in Computer Science*, pp. 226–234, Springer, Amsterdam, The Netherlands, May 2007.

[3] M. Levy and M. Sandler, "A semantic space for music derived from social tags," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR '07)*, pp. 411–416, Vienna, Austria, September 2007.

[4] X. Hu, M. Bay, and S. J. Downie, "Creating a simplified music mood classification ground-truth set," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR '07)*, pp. 309–310, Vienna, Austria, September 2007.

[5] A. Bär, A. Berger, S. Egger, and R. Schatz, "A lightweight mobile TV recommender," in *Proceedings of the 6th European Conference on Interactive TV: A Shared Experience (EuroITV '08)*, M. Tscheligi, M. Obrist, and A. Lugmayr, Eds., vol. 5066 of *Lecture Notes in Computer Science*, pp. 143–147, Springer, Salzburg, Austria, July 2008.

[6] J. Wang, L. Duan, L. Xu, H. Lu, and J. S. Jin, "TV ad video categorization with probabilistic latent concept learning," in *Proceedings of the 9th ACM SIG Multimedia International Workshop on Multimedia Information Retrieval (MIR '07)*, pp. 217–226, Bavaria, Germany, September 2007.

[7] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[8] K. Skreiner, "In the news: machine learning takes on the brain," *IEEE Intelligent Systems*, vol. 23, no. 3, pp. 7–8, 2008.

[9] T. M. Mitchell, S. V. Shinkareva, A. Carlson, et al., "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[10] D. J. Levitin and V. Menon, "Musical structure is processed in "language" areas of the brain: a possible role for Brodmann Area 47 in temporal coherence," *NeuroImage*, vol. 20, no. 4, pp. 2142–2152, 2003.

[11] S. Koelsch and W. A. Siebel, "Towards a neural basis of music perception," *Trends in Cognitive Sciences*, vol. 9, no. 12, pp. 578–584, 2005.

[12] N. Steinbeis and S. Koelsch, "Shared neural resources between music and language indicate semantic processing of musical tension-resolution patterns," *Cerebral Cortex*, vol. 18, no. 5, pp. 1169–1178, 2008.

[13] L. R. Slevc, J. C. Rosenberg, and A. D. Patel, "Language, music and modularity, evidence for shared processing of linguistic and musical syntax," in *Proceedings of the 10th International Conference on Music Perception & Cognition (ICMPC '08)*, Sapporo, Japan, August 2008.

[14] D. Schön, R. L. Gordon, and M. Besson, "Musical and linguistic processing in song perception," *Annals of the New York Academy of Sciences*, vol. 1060, pp. 71–81, 2005.

[15] V. Gallese, "Embodied simulation: from neurons to phenomenal experience," *Phenomenology and the Cognitive Sciences*, vol. 4, no. 1, pp. 23–48, 2005.

[16] V. Gallese and G. Lakoff, "The brain's concepts: the role of the sensory-motor system in conceptual knowledge," *Cognitive Neuropsychology*, vol. 22, no. 3-4, pp. 455–479, 2005.

[17] I. Molnar-Szakacs and K. Overie, "Music and mirror neurons: from motion to 'e' motion," *Social Cognitive and Affective Neuroscience*, vol. 1, no. 33, pp. 235–241, 2006.

[18] L. B. Meyer, "Meaning in music and information theory," *Journal of Aesthetics and Art Criticism*, vol. 15, no. 7, pp. 412–424, 1957.

[19] D. Temperley, *Music and Probability*, MIT Press, Cambridge, Mass, USA, 2007.

[20] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*, MIT Press, Cambridge, Mass, USA, 2006.

[21] R. Jackendoff and F. Lerdahl, "The capacity for music: what is it, and what's special about it?" *Cognition*, vol. 100, no. 1, pp. 33–72, 2006.

[22] C. L. Krumhansl, "Music: a link between cognition and emotion," *Current Directions in Psychological Science*, vol. 11, no. 2, pp. 45–50, 2002.

[23] I. Peretz, R. Gagnon, and S. Hebert, "Singing in the brain: insights from cognitive neuropsychology," *Music Perception*, vol. 21, no. 3, pp. 71–81, 2004.

[24] I. Peretz, M. Radeau, and M. Arguin, "Two-way interactions between music and language: evidence from priming recognition of tune and lyrics in familiar songs," *Memory and Cognition*, vol. 32, no. 1, pp. 142–152, 2004.

[25] M. M. Bradley and P. J. Lang, "Afective norms for English words (ANEW): stimuli, instruction manual and affective ratings," Tech. Rep. C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, Fla, USA, 1999.

[26] E. Pöppel, "A hierarchical model of temporal perception," *Trends in Cognitive Sciences*, vol. 1, no. 2, pp. 56–61, 1997.

[27] W. Kintsch, *Comprehension—A Paradigm for Cognition*, Cambridge University Press, Cambridge, UK, 1998.

Appendix E

Emotional Vectors: Modeling Media from Cognitive Components

Submitted to the *"Journal of Multimedia Systems"* 2008.

Emotional vectors: modeling media from cognitive components

Michael Kai Petersen, Lars Kai Hansen, Andrius Butkus and Martin Schwartz

DTU Informatics, Technical University of Denmark, Building 321,
DK-2800 Kgs.Lyngby, Denmark {mkp,lkh,ab}@imm.dtu.dk

The increasing amounts of streaming and downloadable media becoming available in converged digital broadcast and next generation mobile broadband networks will require intelligent interfaces capable of personalizing the selection of content according to user preferences and moods. We propose an approach to automatically generate atmosphere-like metadata from BBC synopsis descriptions, by applying LSA latent semantic analysis to define the degree of similarity between textual program descriptions and emotional tags in a semantic space.

1. INTRODUCTION

Even though we might think of media as an audiovisual stream of consciousness, we often encode the actual sequences of images framed by crashing waves of sound into strings of words. Language allows us to both share our internal representations as mental concepts, as well as categorize distinct states in the continuous ebb and flow of emotions triggered by our interaction with the external world [1]. In neurobiological terms, making sense of our surroundings involves channeling the input from brainstem and subcortical areas associated with sensory processing and emotion, to anterior frontal areas of the brain, which cognitively speaking turn them into meaningful representations [2]. Meaning that even when the affective aspects associated the sensory input are shifted into the background, our thoughts are always infused with emotional value. Taking the processing within the visual pathway as an example this is not a one way road, but involves extensive feedback loops where affective cues extracted from the incoming signal iteratively influences the balance between cognition and perception [3]. As a result emotional value will impact both the top down processes constrained by our attention, as well as the bottom-up selection of neurons, which when firing in synchrony will generate a conscious perception.

So if both low-level features of media and our emotional responses can be encoded in words, we suggest that this might allow us to define a high level cognitive model emulating how we select media based on affective user preferences. In such a model the bottom-up part would resemble cognitive component analysis [4]. Coined as a term to describe aspects of unsupervised clustering of data, the underlying algorithms approximate how our brain discovers self-organizing patterns when assembling images from lines and edges of visual objects [5], reconstructs words from the statistical regularities of phonemes in speech [6] or learn the meaning of words based on their co-occurrence within multiple contexts [7]. But equally important: cognitive processes involve a large amount of top-down feedback which sculpts the receptive responses of neurons on every level and vastly outnumbers the sensory inputs [8-10]. That is, the brain applies an 'analysis-by-synthesis' approach, which combines a top-down capability to infer structure from bottom-up processing of statistical regularities in what we perceive. A way to emulate this approach of the human brain in relation to search of media, would be to apply unsupervised learning of features based on latent semantics, extracted from synopses texts associated with TV programs or lyrics associated with songs. And combine the bottom-up extracted

representation with top-down aspects of attention reflecting preferred emotional structures, similar to the combinations of user generated affective terms found in tag clouds in social networks like *last.fm*.

We outline in the following sections: the tags used for modeling top-down emotional structure, the bottom-up extraction of latent semantics from texts associated with media, a comparative analysis of emotional patterns retrieved using two different text corpora, followed by a discussion of the potential in combining latent semantics and emotional components to enable personalized search of media.

2. AFFECTIVE DIMENSIONS

If we attempt to model top-down cognitive attentional aspects reflecting affective structure, tag-clouds in music social networks like *last.fm* provide an interesting case. Despite the idiosyncratic character of tags defined by hundred thousands of users, recent studies within music information retrieval have revealed that *last.fm* users often tend to agree on the emotional terms they apply to music. The affective terms which are frequently chosen as tags by users to describe the emotional terms they apply to music. The affective terms which are frequently chosen as tags by users to describe the emotional context of songs seem to form clusters around primary moods like mellow, sad, or more agitated feelings like angry and happy. This correlation between social network tags and the specific music tracks they are associated with, has been used in the music information retrieval community to define a simplified mood ground-truth, reflecting not just the words people frequently use when describing the perceived emotional context, but also which tracks they agree on attaching these tags to [11-12]. Selecting twelve of these frequently used tags:

happy, funny, sexy
romantic, soft, mellow, cool
angry, aggressive
dark, melancholy, sad

makes it possible to define an emotional plane as a basis for extracting latent semantics. Drawing on standard psychological parameters for emotional assessment on the basis of user rated values, affective terms are often mapped out along two axes of valence and arousal [13]. Within this emotional plane the dimension of valence describes how pleasant something is along an axis going from positive to negative associated with words like happy or sad, whereas arousal captures the amount of involvement ranging from passive states like mellow and sad to active aspects of excitation as reflected in tags like angry or happy.

How many different parameters it takes to capture the various components in an affective space has been the subject of a number of studies. A model that seems a good fit to how people describe their emotional states, can be described by five underlying latent variables: anger, sadness, disgust, fear and happiness, In such a space the basic emotions are not necessarily grouped according to whether they are being perceived as pleasant or unpleasant, but often occur simultaneously even if they represent contrasting positive and negative aspects of valence [14]. Empirical results for rating of emotional words, also indicate that certain terms e.g. synonyms for happy or anger seem to be based on one category only and are defined as either positive or negative along a single dimension. Whereas other affective terms appear more complex and appear to be combinations of more emotional categories, like despair being perceived as a mixture of sadness and anxiety, or excitement involving aspects of both happiness and surprise [15].

Users at *last.fm* often describe a track as angry but as music is rarely scary or disgusting, tags that describe these feelings are rarely used. In contrast the much larger number of *TV-Anytime* atmosphere genre terms available for describing broadcast content will also capture these aspects by including words like terrifying in its vocabulary. Apart from whether emotions related to fear are included, we have previously found that

the retrieved latent semantics are not dependent on the number of emotional tags being used. Taking advantage of the higher granularity in the *TV-Anytime* atmosphere controlled terms vocabulary, comical elements might not only be described as simply funny but can be further differentiated as humorous, crazy, silly or wacky. However the patterns of valence and arousal reflected in the retrieved latent semantics, remain largely preserved independent of whether only twelve *last.fm* tags or fifty *TV-Anytime* atmosphere terms are applied [16].

3. SEMANTIC NODES

To generate the bottom-up part of how we cognitively extract meaning from strings of texts, LSA latent semantic analysis models comprehension from word occurrences in multiple contexts, analogous to human language acquisition [7]. Words rarely come shrink-wrapped with a definitive meaning but are continuously modified by the context in which they are set. No matter how many examples of word usage for a verb are listed in a dictionary they remain just that: case stories which illustrate how a predicate will map onto a certain value given a specific argument. Replacing any of the surrounding words in the sentence will create yet another instantiation of the proposition, which we might again interpret differently depending on what phrases come before or after in the text. Instead of attempting to define the specific meaning of a word based on how it fits within a particular grammatical phrase structure, LSA latent semantic analysis [17], models the plethora of meanings a word might have by concatenating all the situations in which it appears and represent them as a single vector within in a high dimensional semantic space [18]. Squeezing as many of the syntactic relations and senses of word usage into a single vector, makes it possible to extract statistical properties based on how often a term appears in a large number of paragraphs. And subsequently condense this representation into meaningful semantic relations constructed from an average of the different contexts in which the word is used [7].

Initially a text corpus is constructed which allows for modeling terms as linear combinations of the multiple paragraphs and sentences they occur in. For this article we have compared two different matrices: the TASA (Touchstone Applied Science Associates, Inc.) which is a collection of fiction and non-fiction texts that an american student will have been exposed to when reaching first year of college, and another corpus assembled from tens of thousands of pages of literature, poetry, wikipedia and news articles. Both of these underlying text corpora can be thought of as resembling human memory where numerous episodes combined with lexical knowledge are encoded into strings of text. Spanned by rows of words and columns of documents, the cells of this huge term-document matrix sum up how frequently each word appears in a corresponding paragraph of text. However in a simple co-occurrence matrix any similarities between words like car and vehicle will be lost as each individual term appears only within its own horizontal row. Nor will it be obvious that a word like rock might mean something completely different depending on which of the contextual columns it appears in. The raw matrix counts of how many times a word occurs in different contexts does therefore not by itself provide a model of comprehension, as we would normally expect texts that describe the same topic to share many of the terms that are used, or imagine that words that resemble each other are also applied in a similar fashion. Most of these relations remain hidden within the matrix, because there are tens of thousands of redundant variables in the original term-document matrix obscuring the underlying semantic structure. Reducing the dimensionality of the original matrix using SVD singular value decomposition [17], the number of parameters can be diminished so we can fit synonymous words or group similar documents into a much smaller number of factors that can be represented within a semantic space.

Geometrically speaking, the terms and documents in the condensed matrix derived from the SVD dimensionality reduction, can be interpreted as points in a k dimensional subspace, which enables us to calculate the degree of similarity between texts based on the dot product of their corresponding vectors [15]. But before comparing terms or documents, the entries in the cells of the matrix need to be adjusted so they reflect how we cognitively perceive associative processes. First by replacing the raw count of how often a word appears in a text by the logarithm of that number. This will smooth the word frequency so it resembles the shape of learning curves typically found in empirical psychological conditioning experiments. Likewise the degree of association of two words both occurring in two documents will be higher than if they each appear twice separately in a text. Here a local weighting function defines how salient the word occurrence is in the corresponding document, and a global weighting function how significant its appearance is among all the contexts [19]. As a next step the word count is divided by the entropy of the term, to ensure that the term frequency will be modified by how much information the word actually adds about the context it appears in. This log-entropy weighting significantly improves the results when compared to a raw word frequency count [7].

Another way to interpret the relations between words forming a semantic neighborhood would be to think of them as nodes constituting a neural network. In such a network of nodes, resembling populations of neurons in our brains, LSA could model the strength of the links connecting one word to another. When we come across a word like `sad' in a phrase, it will create a node in our short term episodic memory, which will in turn trigger neighboring nodes representing words or events that invoke similar connotations in our past memories. The strength of the connections initially based on word co-occurrence are gradually transformed into semantic relations as the links between nodes are being constrained by the limitations of our memory [20]. As a result only those nodes which remain sufficiently activated when our attention shifts towards the next phrase will be integrated into the patterns forming our working memory. And whether these connections grow sufficiently strong for the nodes to reach a threshold level of activation necessary for being integrated in working memory, can be seen as a function of the cosine between the word vectors [21].

When we compare two terms in the LSA semantic space based on the the cosine of the angle between their vectors, values in-between 0.05 and 1 will indicate increasingly significant degrees of similarity between the words, while a negative or low value around 0 will indicate a random lack of correlation. If we for instance select the affective term ‘sad’ and calculate the cosine between the angle of its vector representation and any other word in the text corpus, we can determine which other term vectors are semantically close, and in decreasing order list to what degree they share aspects reflecting the meaning of that word:

1.000000238418579	sad
0.7382655739784241	grief
0.7253139615058899	sorrow
0.6309483647346497	mourn
0.6180344223976135	sigh
0.580369770526886	weep
0.5282069444656372	tear
0.5055677890777588	griev
0.5036925077438354	piti
0.49321797490119934	ala

Looking at these nearest neighbors it would seem that instead of interpreting ‘sad’ isolated as a single vector made from the various documents in which it appears, we might rather think of the meaning of that word as a semantic neighborhood of vectors. In this part of our LSA semantic space these nearest neighbors form a network of nodes, where each word add different aspects to the meaning depending on the strength of their associative links to ‘sad’. So if we imagine text comprehension as a process that combines the words which shape a sentence with the

associations they trigger we can model this as a bottom-up spreading activation process [21]. In this network the strength of links between nodes will be defined by their weights and consequently the connections among all nodes can be mapped out in a connectivity matrix. Being exposed to an incoming word the stimulus will spread from the node generated in episodic memory, to its semantically nearest neighbors stored in long term working memory. How many of these connections grow sufficiently strong for the nodes to be integrated in long term working memory, determines whether our comprehension is reduced to an assembly line where separate words are merely glued together based on the incoming text alone. Or it will instead provide a blueprint for reconstructing a situation model, resembling an animated pin-ball machine where the associations triggered by the words bounce off walls forming an intricate maze of memories. And once reality kicks in, in terms of the constraints posed by the limited capacity of our working memory, what nodes will remain activated could be understood as proportional to the LSA cosine similarity of vectors, triggered by the words being parsed and their nearest neighbors already residing in our memories [22].

4. EMOTIONAL VECTORS

Emulating the 'analysis-by-synthesis' approach of the brain, which combines bottom-up sensory input with top-down preferences for a specific structure, we could use LSA to provide a simplified model of the processes involved in contextual search of media. In this model the bottom-up learning of features would be translated into representing the relations between words and paragraphs as vectors in a semantic space. And top-down, projecting the *last.fm* tags into the semantic space as emotional attractors would make it possible to generate patterns defining to what degree these affective components are reflected in the underlying semantic structure of texts describing media. And taking a selection of short BBC program descriptions as input, we thus compute the cosine similarities between a vector representing the synopsis text against each of the twelve vectors corresponding to the most frequently used *last.fm* emotional tags.

As previously mentioned we have for this experiment used two different text corpora to generate the underlying semantic relations between words defining the basis for the LSA analysis. Using the LSA web-based service available at University of Colorado (www.lsa.colorado.edu), makes it possible to perform the analysis based on the TASA collection of fiction and non-fiction texts, which corresponds to the material an american student has read up to first year of college. It contains 92409 words selected from 119627 paragraphs of primarily language arts, social studies and science texts. We have in parallel run an LSA analysis of the same texts using our own experimental setup, where we have constructed a text corpus consisting of 22829 terms found in 67351 paragraphs, assembled from the poetry and literature volumes forming Harvard Classics, Wikipedia pages on music as well as news articles selected from Reuters (HAWIR).

An analysis of the program 'News night', based on the short description: *News in depth investigation and analysis of the stories behind the day's headline*, based on the TASA text corpus (Fig.1) triggers the emotional tags 'funny' and 'sexy' which might not immediately seem a fitting description, probably caused by these affective terms being directly correlated with the occurrence of the words stories and news within the synopsis. These tags are not triggered when the underlying semantic relations are generated from the HAWIR collection of texts (Fig.2), where instead 'mellow' is activated. The semantic relations between the words might be interpreted not only based on the positive cosine values, but also the negative correlation gives an indication of which emotional tags are not at all triggered by the synopsis text.

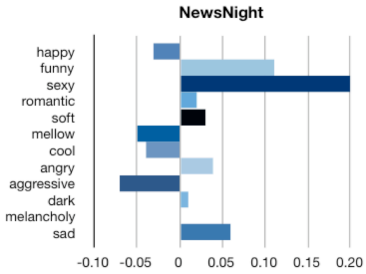


Fig.1, tags triggered by synopsis (TASA).

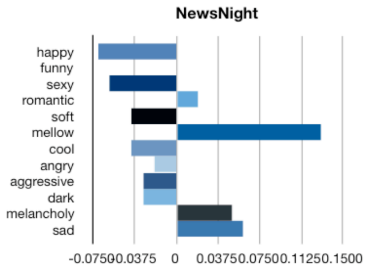


Fig.2, tags triggered by synopsis (HAWIR).

The less serious atmosphere of the lifestyle program 'Ready Steady Cook!' might be somewhat better captures in the synopsis: *Peter Davidson and Bill Ward challenge celebrity chefs to create mouth watering meals in minutes*, which triggers the tag 'romantic' as associated with meals based on the TASA (Fig.3). Whereas the aspects of funny and cool are the ones that are reflected in the HAWIR text corpus (Fig.4)

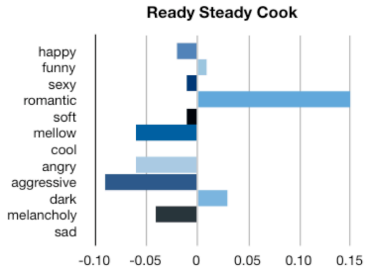


Fig.3, tags triggered by synopsis (TASA).

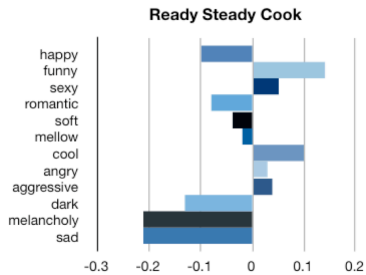


Fig.4, tags triggered by synopsis (HAWIR).

Emotions from the other end of the spectrum are found in the documentary 'I am a boy anorexic', summed up in the synopsis: *Documentary following three youngsters struggling to overcome their obsessive relationship with food as they recover inside a London clinic and then return to the outside world*, which triggers the affective term 'dark' based on the TASA (Fig.5) and 'aggressive' when using the HAWIR text corpus (Fig.6).

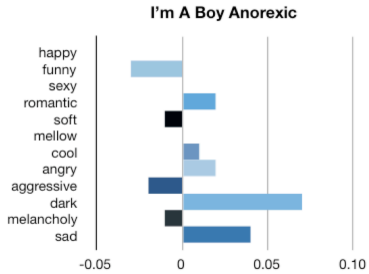


Fig.5, tags triggered by synopsis (TASA).

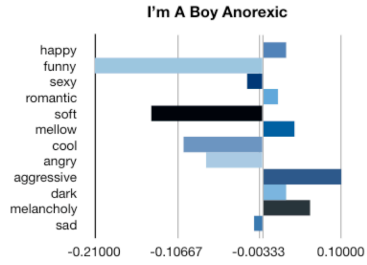


Fig.6, tags triggered by synopsis (HAWIR).

A broader emotional spectrum seems to be reflected in the lifestyle program 'The flying gardener' described by the text: *The flying gardener Chris travels around by helicopter on a mission to find Britain's most inspirational gardens. He helps a Devon couple create a beautiful spring woodland garden. Chris visits impressive local gardens for ideas and reveals breathtaking views of Cornwall from the air*. The synopsis triggers a concentration of passive pleasant elements in the TASA related to the words 'soft', 'mellow' combined with 'happy'. In this context also the tag 'cool' comes out as it has a strong association to the word air contained in the synopsis, while the activation of the tag aggressive appears less explainable (Fig.7). Also in the HAWIR corpus 'soft' comes out, but here coupled with 'happy' and 'melancholy' components (Fig.8)

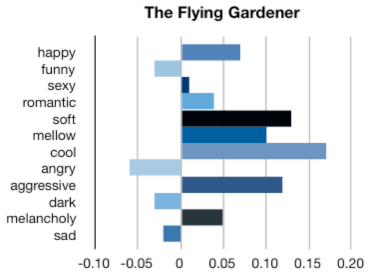


Fig.7, tags triggered by synopsis (TASA).

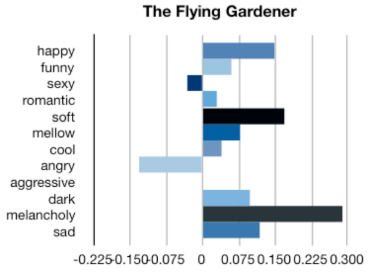


Fig.8, tags triggered by synopsis (HAWIR).

These predominantly positive elements are lacking in the program 'Super Vets' which instead evokes a strong emotional contrast from the text: *At the Royal Vet College Louis the dog needs emergency surgery after a life threatening bleed in his chest and the vets need to find out what is causing the cat Blueboy fits*, where both pleasant and unpleasant active terms like 'happy' and 'sad' stand out in combination with strong emotions reflected by the tag 'romantic' based on the TASA (Fig.9). 'Romantic' and 'sad' are likewise triggered based on the HAWIR text corpus but here complemented by 'funny' (Fig.10). As can be seen from programs like 'The flying gardener' and 'Super Vets' the correlation between the synopsis and the emotional tags might often trigger both combinations of complementary elements as well as contrasting emotional components rather than a monochrome cluster of feelings.

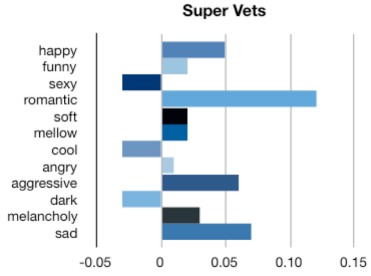


Fig.9, tags triggered by synopsis (TASA).

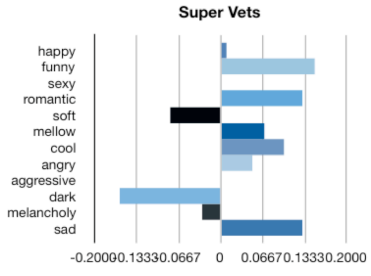


Fig.10, tags triggered by synopsis (HAWIR).

The cult series 'Buffy the vampire slayer' is summed up in the resume: *Shocked from her shallow lifestyle high school cheerleader Buffy learns she is supposed to be a fearsome warrior in the ongoing battle against the bloodsucker's who plague the world*, which based on the

TASA triggers ‘sexy’, ‘aggressive’ and ‘dark’ (Fig. 11), while the ‘cool’ and ‘funny’ aspects become more dominant when based on the HAWIR corpus (Fig.12).

We proceeded to explore whether we could sum up a distinct pattern reflecting an emotional profile pertaining to a TV series, by accumulating the LSA values of correlation between more synopsis texts and emotional tags over several episodes.

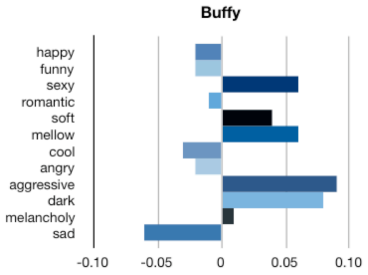


Fig.11, tags triggered by synopsis (TASA).

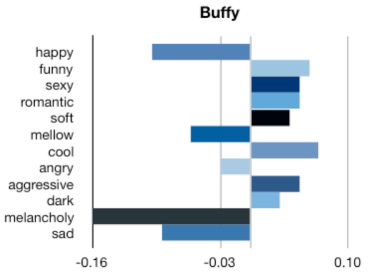


Fig.12, tags triggered by synopsis (HAWIR).

For this purpose we chose the soap ‘East Enders’ and the comedy ‘Two pints of lager’ and analyzed descriptions of six consecutive episodes from each series. When comparing the accumulated LSA correlation between synopsis and affective terms over six episodes of the soap ‘East Enders’, a the tags ‘angry’, ‘happy’ and ‘sad’ stand out both based on the TASA (Fig.13) and HAWIR (Fig.14) corpora.

Whereas the comedy ‘Two pints of lager’ lacks the ‘sad’ component in both text corpora. Instead the affective terms ‘aggressive’ ‘happy’ and ‘funny’ are lightly triggered in the TASA (Fig.15), while ‘funny’ ‘mellow’ and ‘angry’ are more strongly activated based on the HAWIR corpus (Fig.16). So whereas the soap opera appears to cover a wider range of feelings ranging from ‘sad’ to ‘happy’ components, the emphasis in the comedy seems to be shifted towards predominantly happy and funny elements.

These patterns become more clear when instead plotting the emotional components over time in each episode of the soap and comedy respectively. Based on the TASA corpus the distribution in ‘East Enders’ appears as much more dense and emotionally saturated reflecting aspects of ‘arousal’, while the character of ‘Two pints of lager’ seems mirrored in a pronounced clustering of lighter elements of positive valence and an overall sparsity of excitation within the matrix (Fig.17).

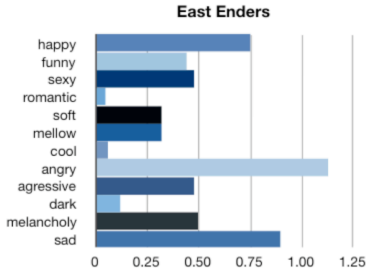


Fig.13 LSA accumulated values (TASA).

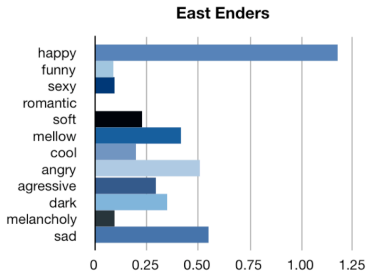


Fig.14, LSA accumulated values (HAWIR).

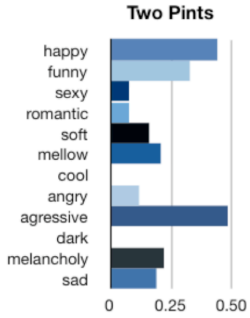


Fig.15, LSA accumulated values (TASA).

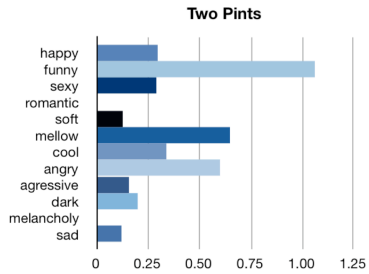


Fig.16, LSA accumulated values (HAWIR).

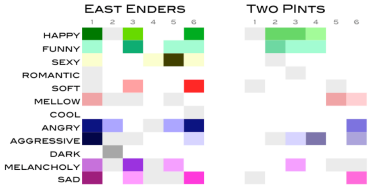


Fig.17, LSA values over 6 episodes (TASA).

Making the same comparison based on the HAWIR corpus the matrices now appear roughly equally saturated. Instead the distinction emerges from the higher separation of components, where the comedy lacks the bottom-heavy 'sad' part in contrast to the soap, while 'funny' is being continuously triggered (Fig.18)

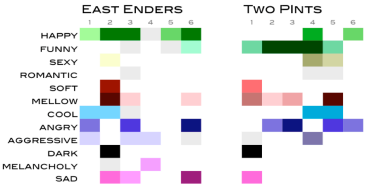


Fig.18, LSA values over 6 episodes (HAWIR).

4. DISCUSSION

If LSA as a model can be interpreted as reflecting a human-like comprehension of language, it has earlier been suggested that one approach to validate its performance is to test how well it understands the meanings of synonymous words [7]. Or in other words measure its ability to model the similarity of terms based on how closely their vector representations are positioned in an LSA semantic space. We therefore submitted our HAWIR corpus based LSA setup to a TOEFL english as a foreign language synonymy test, similar to the one a non-native student will be required to take before entering college. In this test a possible question would be to decide the meaning of an adjective like *frightened,* as in *quivering*, and the task is then to pick the right word among four suggested alternatives: *tremulous*, *craven*, *succulent* or *congenial*. Projecting first *frightened* plus the possible answers into the semantic space, followed by *quivering* together with the four alternatives, we get two LSA correlation values for each possible answer. We then use the sum of these retrieved values, e.g. *frightened*, *tremulous;* 0.09 + *quivering*, *tremulous;* 0.61, for each of the four choices A,B,C and D to determine which of the answers have the highest correlation to the question. Subsequently we check whether the synonym that has the highest cosine correlation is also the correct answer to the question.

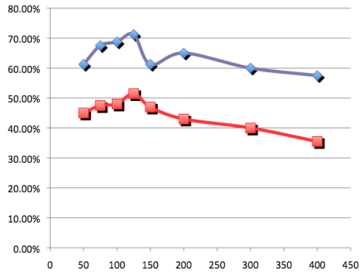


Fig.19, Percentage of correct answers to TOEFL test based on HAWIR corpus, which attains maximum when using 125 factors for the SVD dimensionality reduction in the LSA analysis.

Testing the language skills of our HAWIR text corpus based LSA model with eighty TOEFL questions we get 71,25 % correct answers, when the correlation is based on both the query and its context e.g. the sum of values from *frightened* and *quivering*, or 51,50 % when considering the two query terms separately. This might be compared to previously reported TOEFL synonymy test results achieving 64,4 % correct answers, based on an LSA text corpus consisting of articles from the *Groliers Academic American Encyclopedia*, which in turn appears to be equal to the average percentage of correct answers obtained by non-English speaking applicants [7]. As can be seen from the graph (Fig.19) the percentage of correct answers depends on the number of dimensions chosen for the SVD decomposition of the original term-document matrix. In the case of our HAWIR text corpus, we achieve the best fit for representing synonymous words or grouping similar documents within an LSA semantic space when the dimensionality of the condensed matrix is reduced to 125 factors.

Analyzing the patterns of correlation between affective words and synopses plotted over time for the soap and comedy based on the TASA (Fig.17) or HAWIR (Fig. 18) text corpora, their difference might to a certain extent be explained by considering what terms are the nearest neighbors to the emotional tags in the respective LSA semantic spaces. Taking 'sad' as an example the words:

0.58 happy
0.54 feelings
0.53 feel

are the closest positioned vectors in the TASA corpus. And similar for the emotional tag 'happy' the words:

0.59 sad
0.49 loved
0.42 unhappy

make up the immediate semantic neighborhood. So both 'sad' and 'happy' seem to be defined by their respective antonyms. It might not be an unreasonable assumption that we define 'happy' as being 'not sad', but if the lexical opposites come out as the nearest neighbors, this would explain why the TASA based patterns (Fig.17) shows a tendency to consistently activate emotional tags from both ends of the spectrum.

In contrast the nearest neighbors to 'sad' in the LSA semantic space based on the HAWIR corpus (Fig.18) are:

0.74 grief
0.73 sorrow
0.63 mourn

meaning that the antonym is here not used to define the term, nor in the case of 'happy'

0.54 joy
0.48 enjoy
0.42 bliss

is the lexical opposite found among the top ten closest neighbors. This appears to result in a more clear separation of the affective terms in the HAWIR corpus, which might here also explain the ability to distinguish between 'happy' and 'funny', as opposed to the TASA based patterns where these emotional tags seems to color bleed into each other.

Comparing the emotional components retrieved from the LSA analysis of the

synopses of 'East Enders' and 'Two Pints' against the actual labels they have been manually annotated with in the BBC metadata, using *TV-Anytime* genre atmosphere terms, they seem to be largely in agreement. The comedy 'Two Pints' has been indexed as 'humorous, silly, irreverent, fun, wacky, crazy' and one might argue that most of these components also come out in the LSA analysis. In the case of the soap 'East Enders' the episodes have been annotated as 'gripping, gritty, gutsy', and these aspects might be interpreted as reflected in the stark accumulated contrasts spanning a range of contrasting emotions in the LSA analysis.

5. CONCLUSION

Projecting BBC synopsis descriptions into an LSA space using *last.fm* tags as emotional buoys, we have demonstrated an ability to extract patterns reflecting combinations of emotional components. Analyzing the emotional components reflected in the synopsis descriptions over a sequence of episodes, we have been able to separate these aspects into patterns which might in the future be used as a basis for searching similar media based on the characteristics of sparsity and overall distribution of emotional components. While each synopsis triggers an individual emotional response related to a specific episode, general patterns still emerge when accumulating the LSA correlation between synopsis and emotional tags over consecutive episodes, which enables us to differentiate between a comedy and a soap based on a textual description alone. We therefore propose that emotional components describing the content of media might be retrieved as latent semantics by using affective terms as sensors in a semantic space, and we suggest that LSA might be applied to extract structural patterns from synopsis descriptions as a basis for automatically generating mood-based recommendations. Though the synopsis descriptions trigger both combinations of complementary elements as well as contrasting emotional components rather than a monochrome affective response, they nevertheless pertain to distinct patterns which we suggest might be used as a basis to build emotional patterns capturing user preferences.

6. REFERENCES

[1] L.F. Barrett: *Solving the emotion paradox: categorization and experience of emotion*, Personality and social psychology review (10:1), pp.20-46, 2006

[2] S. Duncan and L.F. Barrett: *Affect is a form of cognition: a neurobiological analysis*, Cognition & Emotion (21:6), pp.1184-1211, 2007

[3] J. Storbeck and G.L. Clore: *On the interdependence of cognition and emotion*, Cognition & Emotion (21:6), pp.1212-1237, 2007

[4] L.K. Hansen and L. Feng: *Cogito componentiter - ergo sum*, In: J. Rocha et al.(eds): Independent Component Analysis and Blind Signal Separation, LNCS 3886, pp.446-453, 2006

[5] A.J.Bell and T.T.Sejnowski: *The independent components of natural scenes are edge filters*, Vision Research (37:23), pp.3327-3338, 1997

[6] L.Feng and L.K.Hansen: *On phonemes as cognitive components of speech*, Proceedings of IAPR workshop on cognitive information processing, 2008

[7] T.K. Landauer and S.T.Dumais: *A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge*, Psychological Review (104:2), pp.211-240, 1997

[8] J. Cudeiro and A.M. Sillito: *Looking back: corticothalamic feedback and early visual processing*, Trends in Neurosciences (29:6), pp.298-306, 2006

[9] A.M. Sillito, J. Cudeiro and H.E. Jones: *Always returning: feedback and sensory processing in visual cortex and thalamus*, Trends in Neurosciences (29:6), pp.307-316, 2006

[10] J.H.R Maunsell and S. Treue: *Feature-based attention in visual cortex*, Trends in Neurosciences (29:6), pp.317-322, 2006

[11] M. Levy and M. Sandler: *A semantic space for music derived from social tags*, Proceedings of the 8th International Conference on Music Information Retrieval, pp.411-416, 2007

[12] X. Hu, M. Bay and S.J.Downie: *Creating a simplified music mood classification ground-truth set*, Proceedings of the 8th International Conference on Music Information Retrieval, pp.309-310, 2007
[13] M. M. Bradley and P.J. Lang: *Affective norms for English words (ANEW), Stiumuli, instruction manual and affective ratings*, The Center for Research in Psychophysiology, University of Florida, 1999
[14] M. J. Power: *The structure of emotion: an empirical comparison of six models*, Cognition & Emotion (20:5), 2006
[15] G. P. Strauss and D.N.Allen: *Emotional intensity and categorisation ratings for emotional nonemotional words*, The Center for Research in Psychophysiology, University of Florida, 1999
[16] M. K. Petersen and A. Butkus: *Extracting moods from songs and BBC programs based on emotional context*, International Journal of Multimedia Brodcasting, 2008
[17] G.W. Furnas, S. Deerwester, S.T. Dumais, T.K. Landauer, R. Harshman, L.A. Streeter and K.E. Lochbaum: *Information retrieval using a singular value decomposition model of latent semantic structure*, In: 11th annual international SIGIR conference, pp.465-480, 1988
[18] W. Kintsch: *Predication*, Cognitive Science (25:2) pp.173-202, 2001
[19] D.I. Martin and M.W. Berry: *Mathematical foundations behind latent semantic analysis*, In: Handbook of latent semantic analysis, Erlbaum, 2007 [20] S. Derwester, S.T. Dumais, G.W. Furnas, W. George, T. Landauer and R. Harshman: *Indexing by latent semantic analysis*, Journal of the American Society for Information Science (41:6), pp.391-407, 1990
[21] W.Kintsch: *Comprehension - a paradigm for cognition*, Cambridge University Press, 1998
[22] W.Kintsch, V.L. Patel and A.K Ericsson: *The role of long-term working memory in text comprehension*, Psychologia (42), pp.186-198, 1999

Appendix F

The Examples of the Emotional Patterns Extracted from the Lyrics

The rest of the *emotional patterns* extracted from the chosen 25 songs are presented here in this appendix.

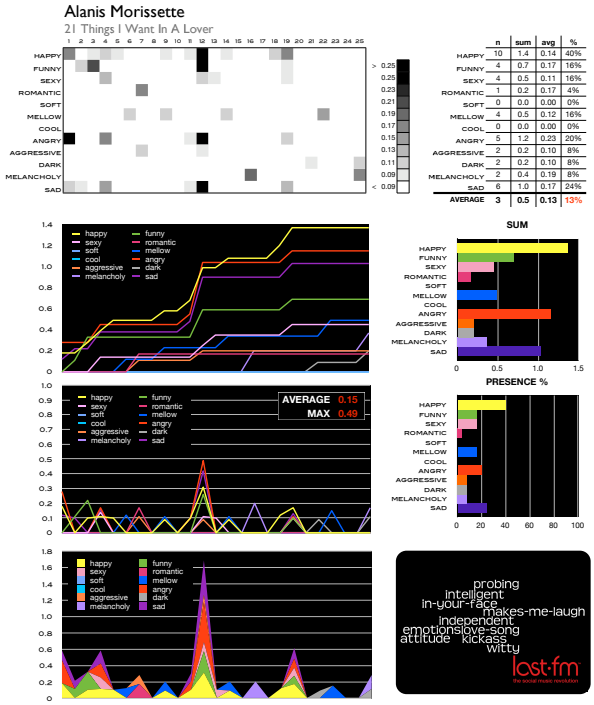


Figure F.1: The extracted emotional patterns from the lyrics of the Alanis Morissette song “21 things i want in a lover”.

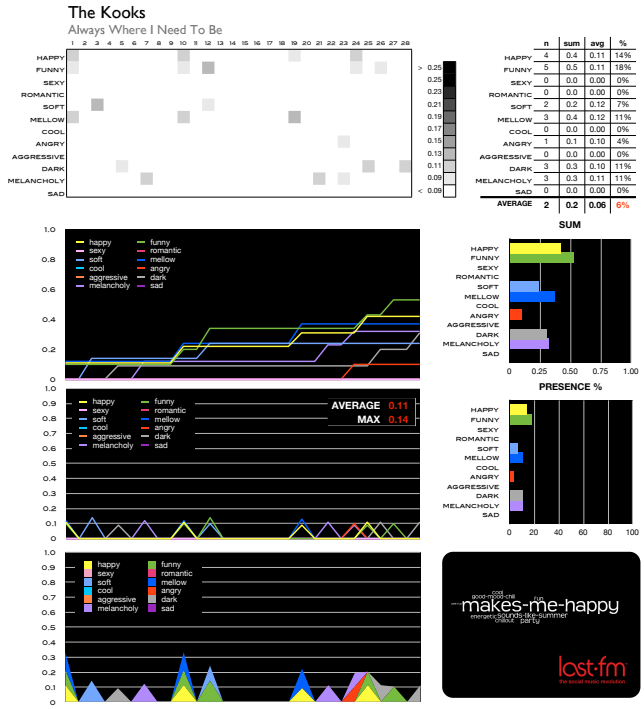


Figure F.2: The extracted emotional patterns from the lyrics of the Kooks song "Always where i need to be".

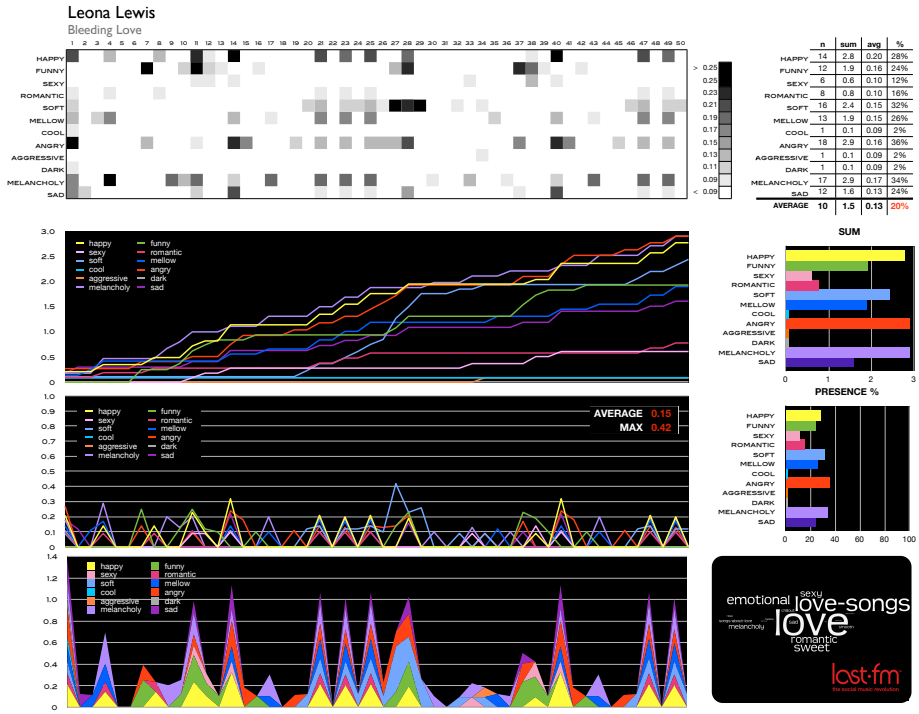


Figure F.3: The extracted emotional patterns from the lyrics of the Leona Lewis song “Bleeding love”.

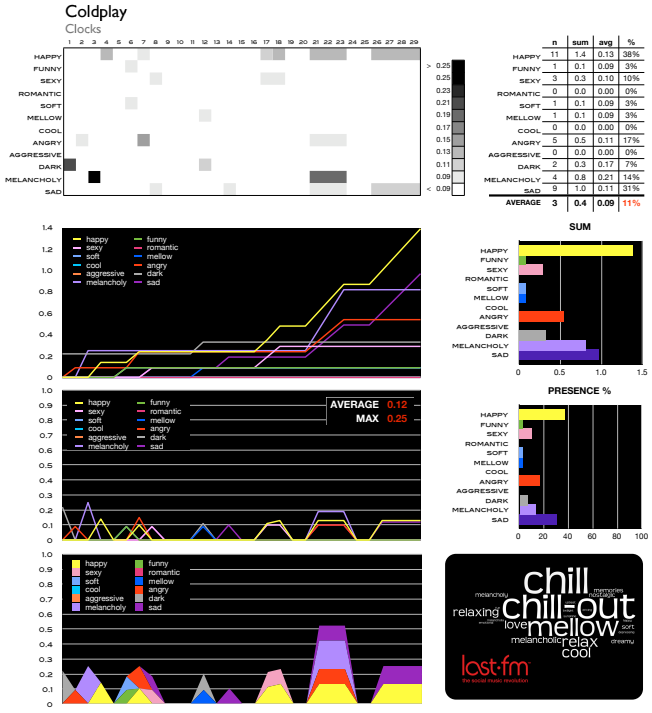


Figure F.4: The extracted emotional patterns from the lyrics of the Coldplay song “Clocks”.

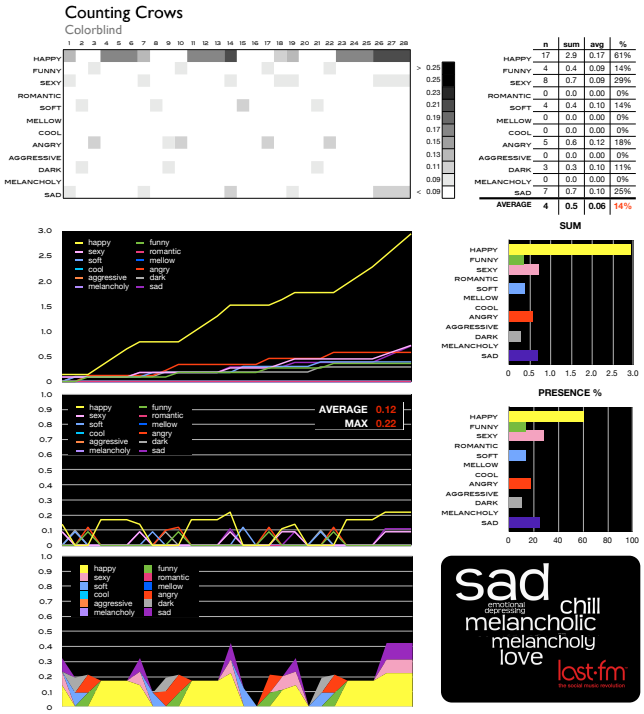


Figure F.5: The extracted emotional patterns from the lyrics of the Counting Crows song “Colorblind”.

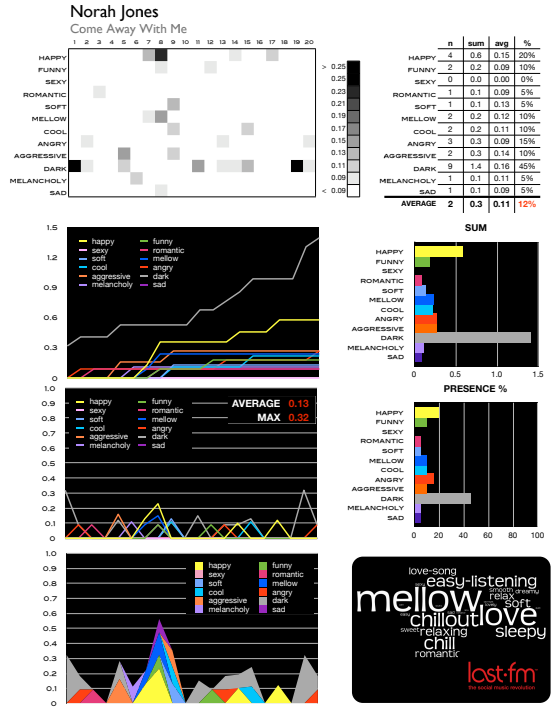


Figure F.6: The extracted emotional patterns from the lyrics of the Norah Jones song “Come away with me”.

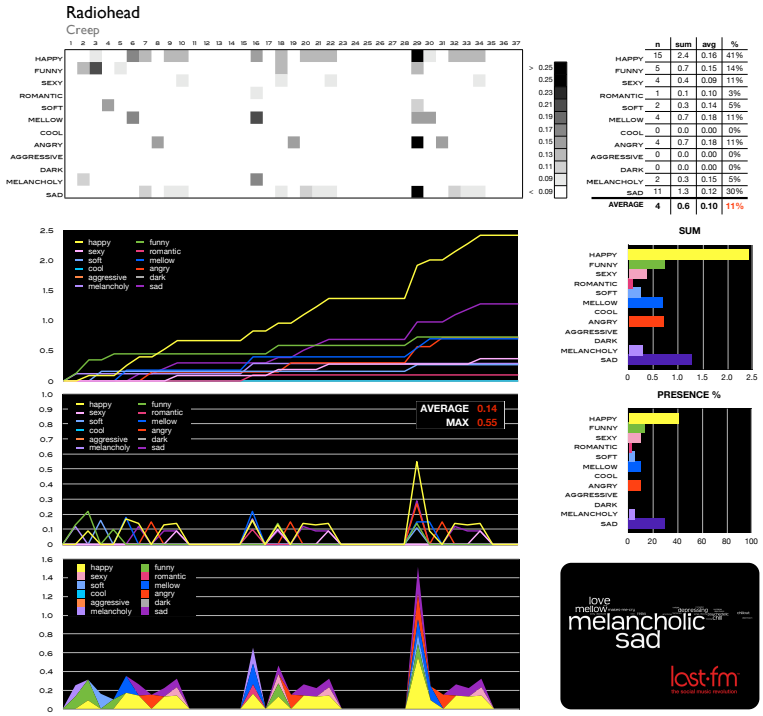


Figure F.7: The extracted emotional patterns from the lyrics of the Radiohead song “Creep”.

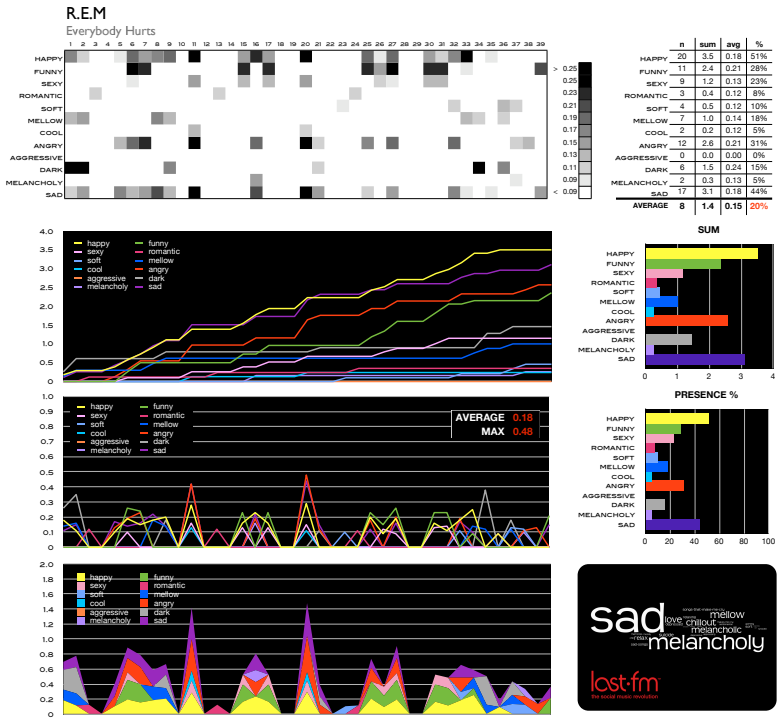


Figure F.8: The extracted emotional patterns from the lyrics of the R.E.M song "Everybody hurts".

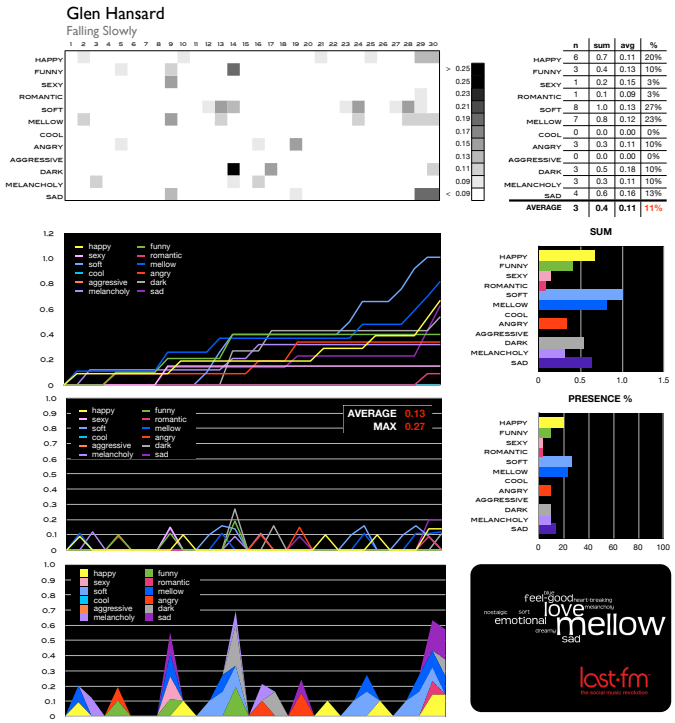


Figure F.9: The extracted emotional patterns from the lyrics of the Glen Hansard song “Falling slowly”.

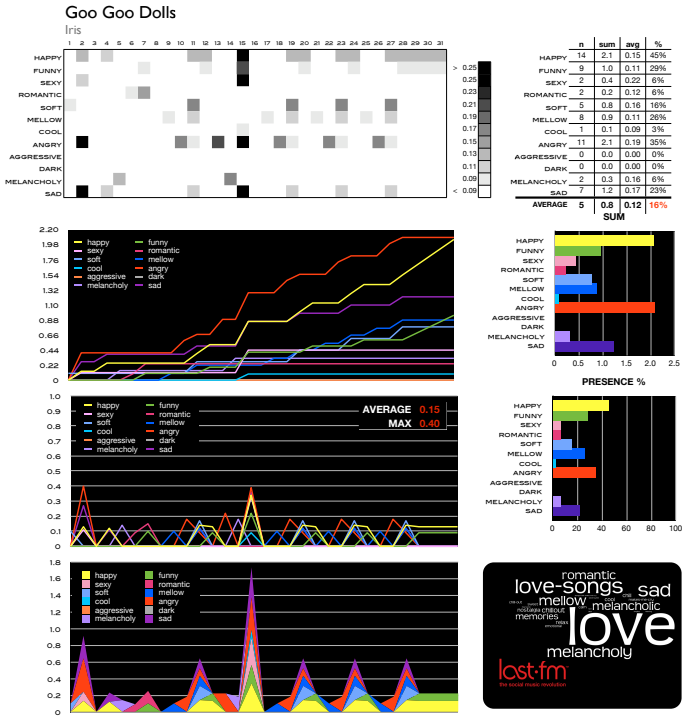


Figure F.10: The extracted emotional patterns from the lyrics of the Goo Goo Dolls song “Iris”.

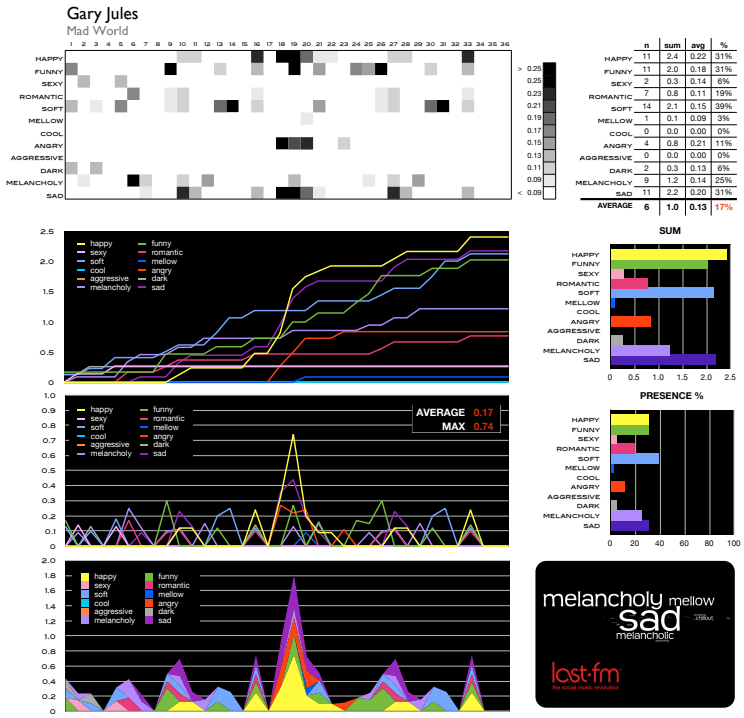


Figure F.11: The extracted emotional patterns from the lyrics of the Gary Jules song “Mad world”.

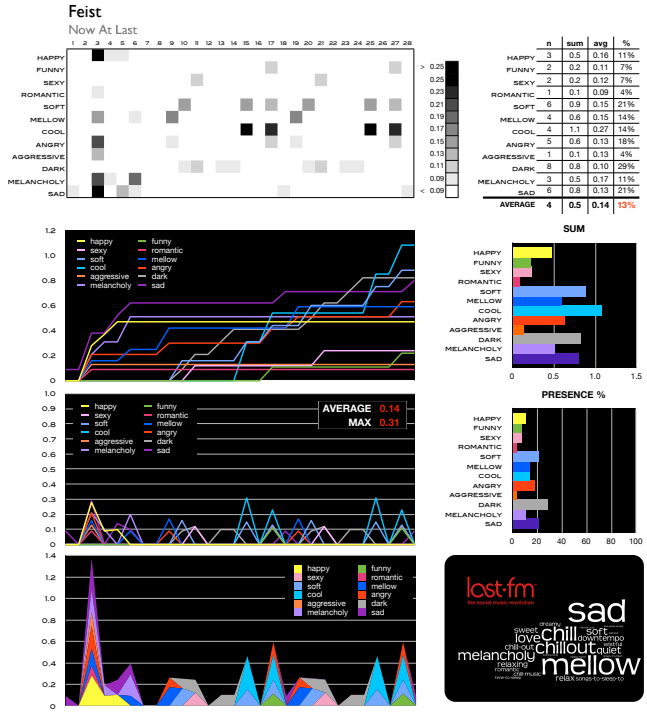


Figure F.12: The extracted emotional patterns from the lyrics of the Feist song "Now at last".

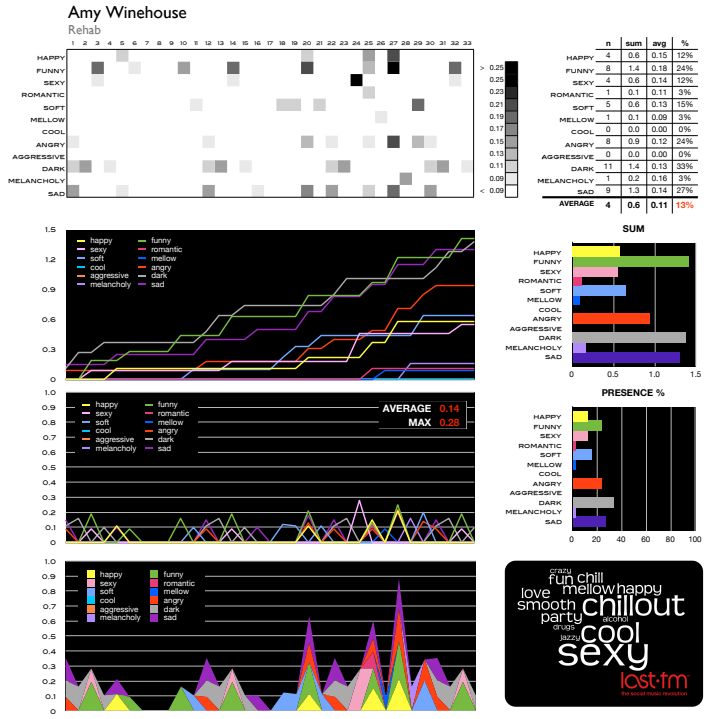


Figure F.13: The extracted emotional patterns from the lyrics of the Amy Winehouse song "Rehab".

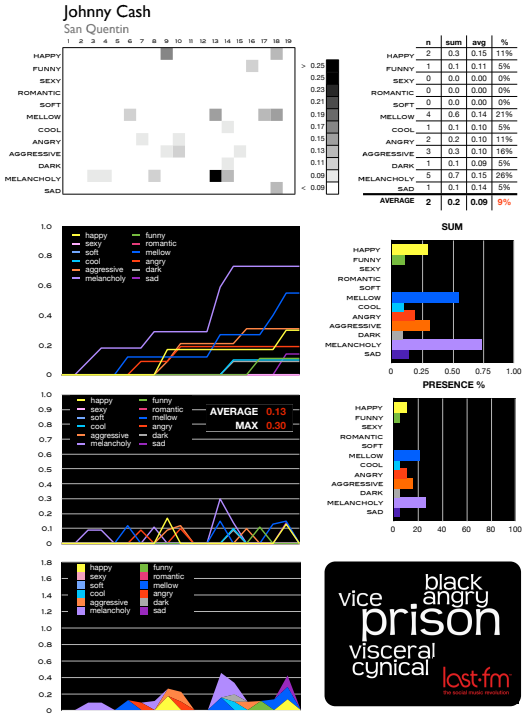


Figure F.14: The extracted emotional patterns from the lyrics of the Johnny Cash song “San Quentin”.

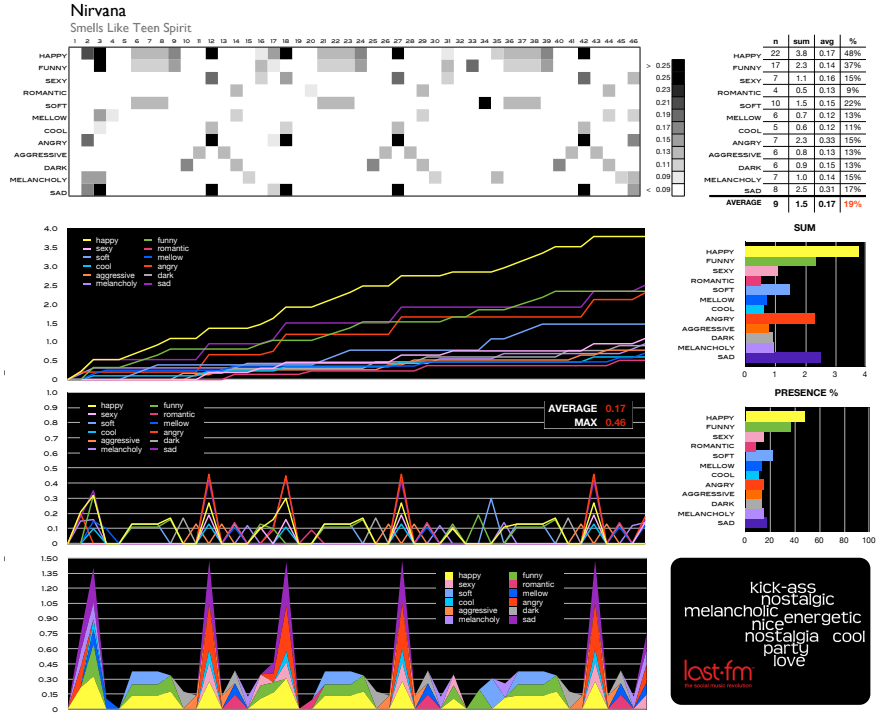


Figure F.15: The extracted emotional patterns from the lyrics of the Nirvana song “Smells like teen spirit”.

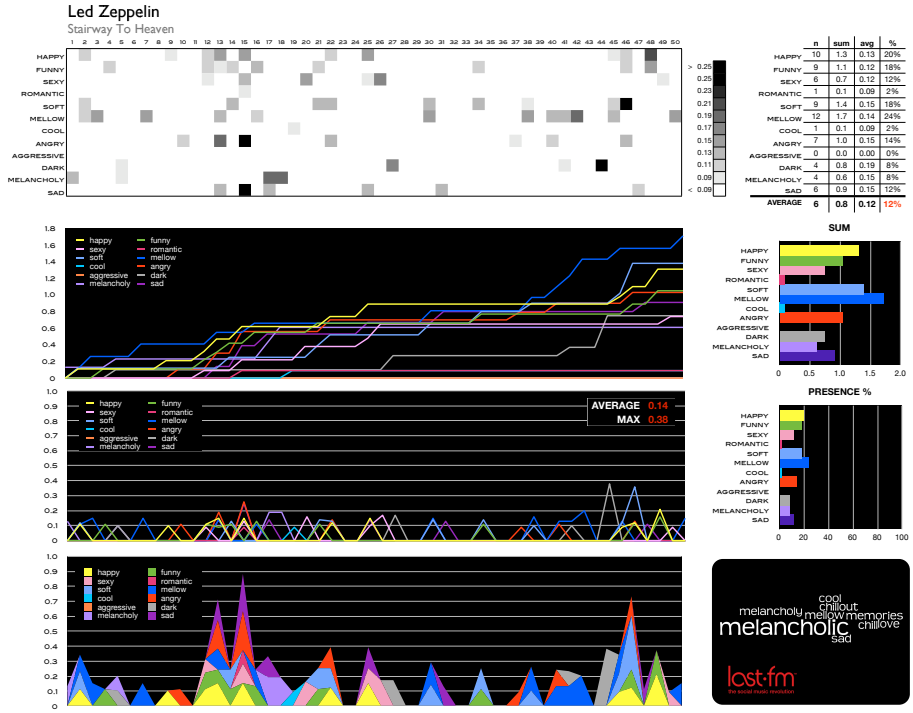


Figure F.16: The extracted emotional patterns from the lyrics of the Led Zeppelin song “Stairway to heaven”.

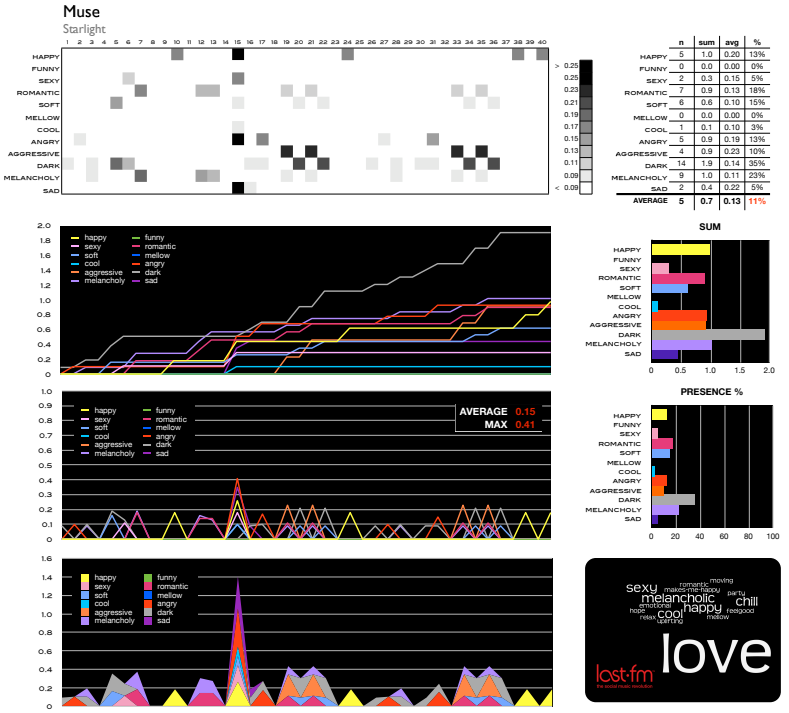


Figure F.17: The extracted emotional patterns from the lyrics of the Muse song "Starlight".

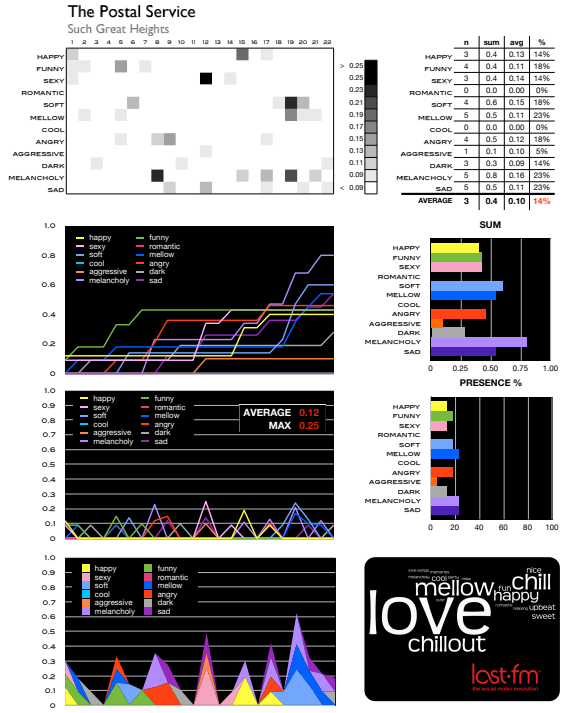


Figure F.18: The extracted emotional patterns from the lyrics of the Postal Service song “Such great heights”.

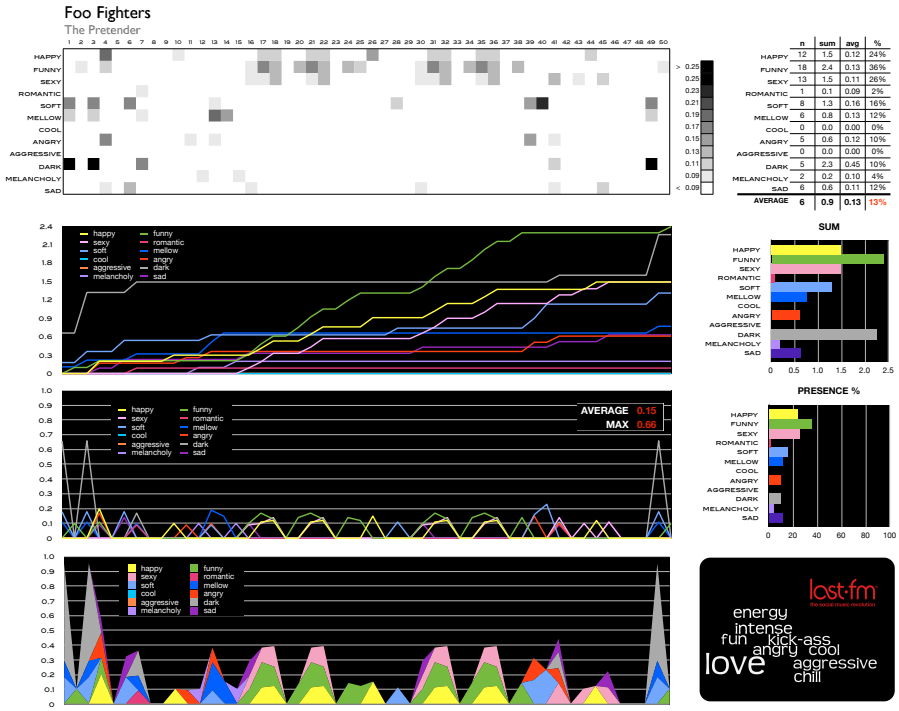


Figure F.19: The extracted emotional patterns from the lyrics of the Foo Fighters song “The pretender”.

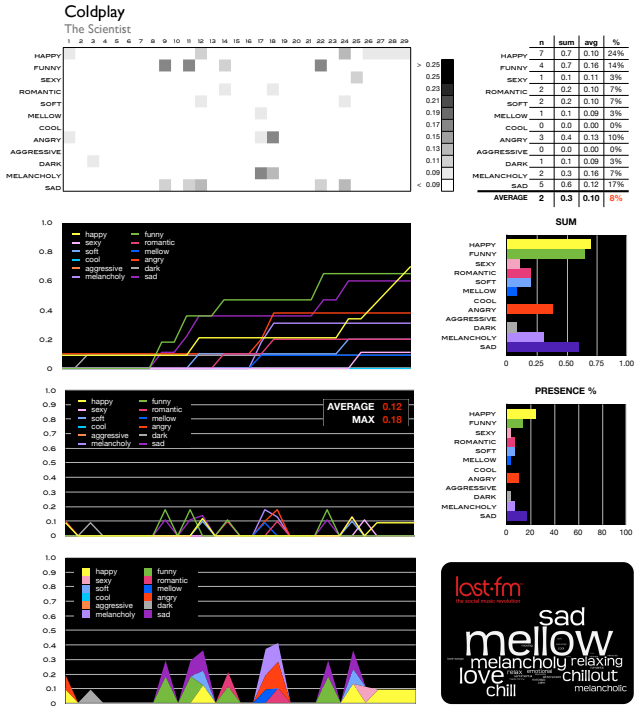


Figure F.20: The extracted emotional patterns from the lyrics of the Coldplay song “The scientist”.

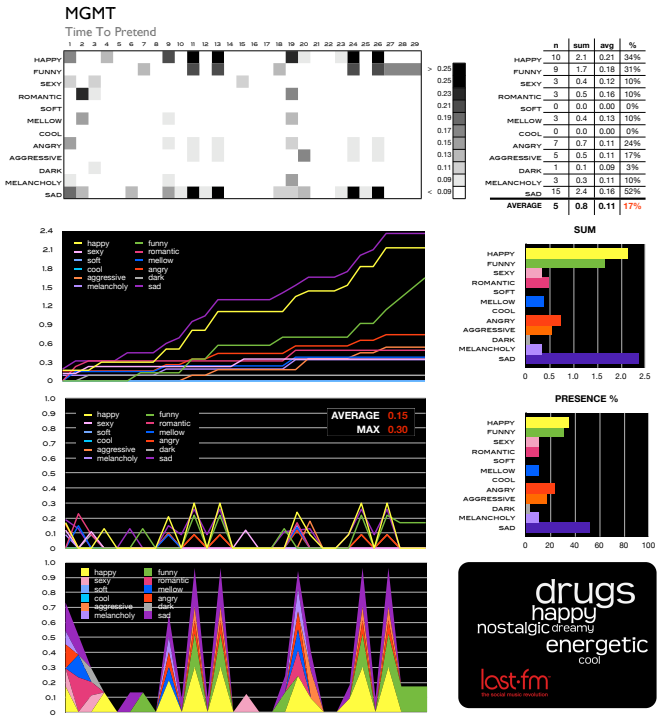


Figure F.21: The extracted emotional patterns from the lyrics of the MGMT song “Time to pretend”.

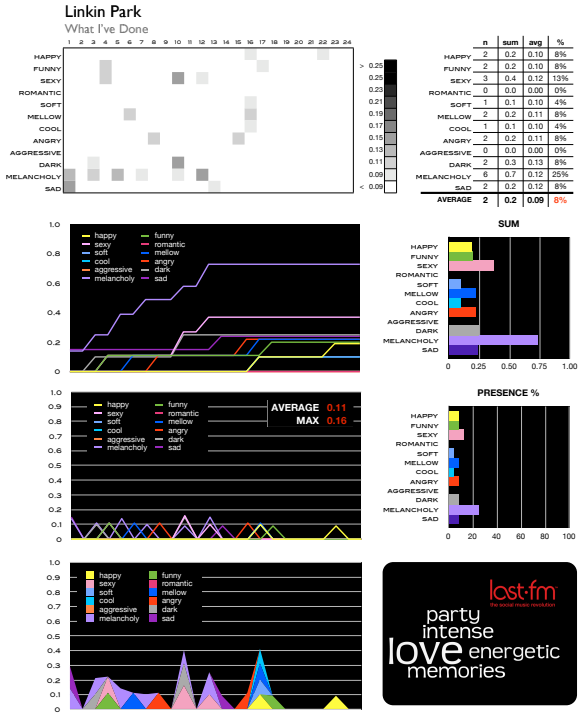


Figure F.22: The extracted emotional patterns from the lyrics of the Linkin Park song “What i’ve done”.

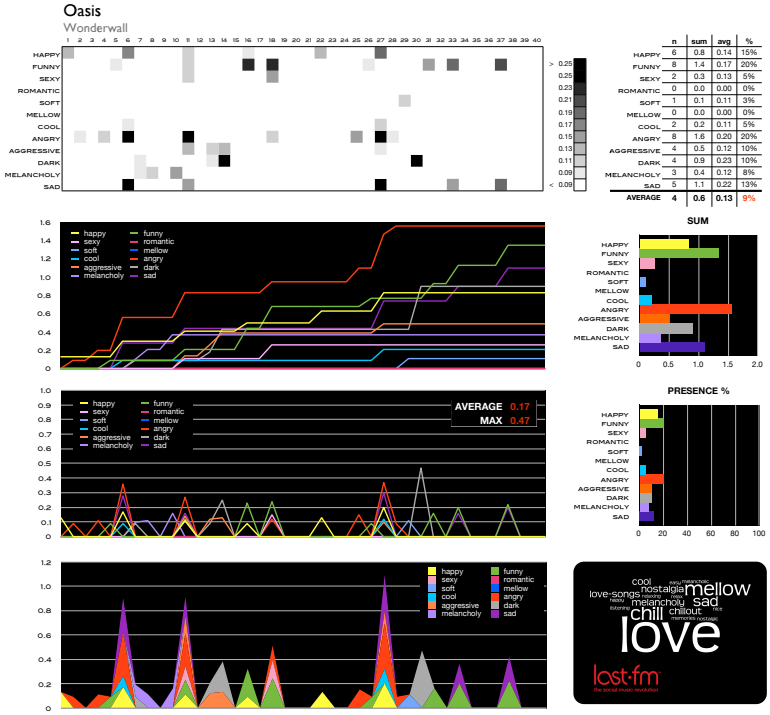


Figure F.23: The extracted emotional patterns from the lyrics of the Oasis song "Wonderwall".

Bibliography

[Adomavicius and Kwon, 2007] Adomavicius, G. and Kwon, Y. (2007). New recommendation techniques for multicriteria rating systems. *IEEE Intelligent Systems*, pages 48–54.

[Adomavicius and Tuzhilin, 2005a] Adomavicius, G. and Tuzhilin, A. (2005a). Personalization technologies: A process-oriented perspective. *Communications of ACM*, 48(10):83–90.

[Adomavicius and Tuzhilin, 2005b] Adomavicius, G. and Tuzhilin, A. (2005b). Toward the next generation of recommender systems- a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

[Akkermans et al., 2006] Akkermans, P., Aroyo, L., and Bellekens, P. (2006). ifanzy: Personalised filtering using semantically enriched tv-anytime content. *proceedings of ESWC*.

[Anderson, 2004] Anderson, C. (2004). The long tail. *Wired*.

[Anderson, 2006] Anderson, C. (2006). *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion.

[Anderson et al., 2003] Anderson, M., Ball, M., Boley, H., Greene, S., Howse, N., Lemire, D., and McGrath, S. (2003). Racofi: A rule-applying collaborative filtering system. *Proceedings of COLA*.

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

[Balabanovic, 1997] Balabanovic, M. (1997). An adaptive web page recommendation service. *Proceedings of the first international conference on Autonomous agents*.

[Balabanovic and Shoham, 1997] Balabanovic, M. and Shoham, Y. (1997). Fab - content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72.

[Basu et al., 1998] Basu, C., Hirsh, H., and Cohein, W. (1998). Recommendation as classification: Using social and content-based information in recommendation. *Proceedings of the Fifteenth National Conference on Artificial Intelligence*.

[Belkin and Croft, 1992] Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, 35(12):29–38.

[Billsus and Pazzani, 1998] Billsus, D. and Pazzani, M. J. (1998). Learning collaborative information filters. *Proceedings of the Fifteenth International Conference on Machine Learning*, 54.

[Bradley and Lang, 1999] Bradley, M. and Lang, P. (1999). Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.

[Bradley et al., 1998] Bradley, P. S., Fayyad, U. M., and Reina, C. A. (1998). Scaling clustering algorithms to large databases. *Knowledge Discovery and Data Mining*, pages 9–15.

[Breese et al., 1998] Breese, J., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *Uncertainty in Artificial Intelligence. Proceedings of the Fourteenth Conference*, pages 43–52.

[Brooks, 1978] Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. *Cognition and categorization*, 3:170–211.

[Brynjolfsson et al., 2007] Brynjolfsson, E., Hu, Y., and Simester, D. (2007). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. working paper.

[Brynjolfsson et al., 2003] Brynjolfsson, E., Hu, Y., and Smith, M. D. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, 11(49):1580–1596.

[Brynjolfsson et al., 2006] Brynjolfsson, E., Hu, Y., and Smith, M. D. (2006). From niches to riches: The anatomy of the long tail. *Sloan Management Review*, 47(4):67–71.

[Burke, 2002] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12:331–370.

[Butkus, 2006] Butkus, A. (2006). Media personalization using tv-anytime phase 2. In *Proceedings of the 3rd International CICT Conference, Mobile and Wireless Content, Services and Networks*.

[Butkus and Petersen, 2007] Butkus, A. and Petersen, M. K. (2007). *Semantic Modelling Using TV-Anytime Metadata in DVB-H mobile broadcast*. Springer Verlag.

[Carson et al., 2002] Carson, C., Belongie, S., Greenspan, H., and Mal, J. (2002). Blobworld: image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Learning*, 24(8).

[Chandler, 2003] Chandler, D. (2003). An introduction to genre theory.

[Collier, 2007] Collier, G. L. (2007). Beyond valence and activity in the emotional connotations of music. *Psychology of Music*.

[Corthaut et al., 2006] Corthaut, N., Govaerts, S., and Duval, E. (2006). Moody tunes: The rockanango project.

[Crowston and Kwasnik, 2003] Crowston, K. and Kwasnik, B. H. (2003). Can document-genre metadata improve information access to large digital collections. *Library Trends*, 52(2):345–361.

[DCMI, 2008] DCMI (2008). Dublin core metadata element set, version 1.1. Technical report, Dublin Core Metadata Initiative.

[Dean, 2005] Dean, K. (2005). Copyrights keep tv shows off dvd. *Wired*.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

[Delgado and Ishii, 1999] Delgado, J. and Ishii, N. (1999). Memory-based weighted-majority prediction. *ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation*.

[Dopkins, 1997] Dopkins, S. (1997). Comparing exemplar and prototype models of categorization. *Canadian Journal of Experimental Psychology*, 51(3):212–230.

[Dronf et al., 2000] Dronf, J., Mitchell, R., Siviter, P., and Boyne, C. (2000). Cofind - an experiment in n-dimensional collaborative filtering. *Journal of Network and Computer Applications*, 23(2):131–142.

[Dumais, 1990] Dumais, S. T. (1990). Enhancing performance in latent semantic indexing (lsi) retrieval. *Bellcore (Bell Communications Research) document TM-ARH-017527*.

[Durrell, 1985] Durrell, W. R. (1985). *Data Administration: A Practical Guide to Data Administration*. McGraw-Hill.

[ETSI, 2003] ETSI (2003). En 300 468 digital video broadcasting (dvb); specification for service information (si) in dvb systems. Technical report, ETSI.

[ETSI, 2004a] ETSI (2004a). Ts 102 822-2 broadcast and on-line services: Search, select, and rightful use of content on personal storage systems ("tv-anytime phase 1"); part 2: System description. Technical report, ETSI.

[ETSI, 2004b] ETSI (2004b). Ts 102 822-3-1 broadcast and on-line services: Search, select, and rightful use of content on personal storage systems ("tv-anytime phase 1"); part 3: Metadata; sub-part 1: Metadata schemas. Technical report, ETSI.

[ETSI, 2006] ETSI (2006). 102 822-3-3 broadcast and on-line services: Search, select, and rightful use of content on personal storage systems ("tv-anytime"); part 3: Metadata; sub-part 3: Phase 2 - extended metadata schema. Technical report, ETSI.

[Faria et al., 2006] Faria, G., Henriksson, J. A., Stare, E., and Tamola, P. (2006). Dvb-h: Digital broadcast services to handheld devices. *IEEE*, 94(1):194–209.

[Furnas et al., 1983] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1983). Statistical semantics: Analysis of the potential performance of key-word information systems. *Bell System Technical Journal*, 62(6):1753–1806.

[Garden and Dudek, 2005] Garden, M. and Dudek, G. (2005). Semantic feedback for hybrid recommendations in recommendz. *Proceedings. The 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service*, pages 754–759.

[Garden and Dudek, 2006] Garden, M. and Dudek, G. (2006). Mixed collaborative and content-based filtering with user-contributed semantic features. *Proceedings of the 21st AAAI National Conference on Artificial Intelligence*.

[Gärdenfors, 2000] Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press.

[Gärdenfors, 2001] Gärdenfors, P. (2001). Concept learning: A geometrical model. In *Proceedings of the Aristotelian Society*, volume 101, pages 163–183.

[Gärdenfors, 2004] Gärdenfors, P. (2004). Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2):9–27.

[Geleijnse et al., 2007] Geleijnse, G., Schedl, M., and Knees, P. (2007). The quest for ground truth in musical artist tagging in the social web era.

[Getoor and Sahami, 1999] Getoor, L. and Sahami, M. (1999). Using probabilistic relational models for collaborative filtering. *Workshop on Web Usage Analysis and User Profiling (WEBKDD'99)*.

[Gevers and Smeulders, 2004] Gevers, T. and Smeulders, A. W. (2004). Content-based image retrieval: An overview. *Emerging Topics in Computer Vision*, 20.

[Goldberg et al., 1992] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70.

[Goldberg et al., 2001] Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151.

[Good et al., 1999] Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J. L., and Riedl, J. (1999). Combining collaborative filtering with personal agents for better recommendations.pdf. *Proceedings of AAAI*.

[Gutta et al., 2000] Gutta, S., Kurapati, K., Lee, K., Martino, J., Milanski, J., Schaffer, J. D., and Zimmerman, J. (2000). Tv content recommender system. *Proceedings of the 17th national confernce on artificial intelligence*, pages 1121–1122.

[Ha, 2006] Ha, S. H. (2006). Digital content recommender on the internet. *IEEE Intelligent Systems*, 21(2):70–77.

[Hendler, 2007] Hendler, J. (2007). The dark side of the semantic web. *IEEE Intelligent Systems*, 22(1):2–4.

[Hevner, 1935] Hevner, K. (1935). The affective character of major an minor modes in music. *American Journal of Psychology*, 47:103–119.

[Hill et al., 1995] Hill, W., Stead, L., Rosenstein, M., and Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 194–201.

[Hillmann, 2005] Hillmann, D. (2005). *Using Dublin Core*. Dublin Core Metadata Initiative.

[Hofmann, 2003] Hofmann, T. (2003). Collaborative filtering via gaussian probabilistic latent semantic analysis. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 259–266.

[Hofmann, 2004] Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):89–115.

[Hornik, 2006] Hornik, D. (2006). Chris anderson strikes again: The economy of abundance.

[Hotho et al., 2006] Hotho, A., Jaschke, R., Schmitz, C., and Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. *The Semantic Web: Research and Applications*, pages 411–426.

[Hu et al., 2007] Hu, X., Bay, M., and Downie, S. (2007). Creating a simplified music mood classification ground-truth set. *Proceedings of the 8th International Conference on Music Information Retrieval, Austrian Computer Society*, pages 309–310.

[Huang et al., 2007] Huang, Z., Zeng, D., and Chen, H. (2007). A comparison of collaborative-filtering recommendation algorithms for e-commerce. *IEEE Intelligent Systems*, pages 68–78.

[Hurley et al., 2007] Hurley, N. J., O'Mahony, M. P., and Silvestre, G. C. (2007). Attacking recommender systems: A cost-benefit analysis. *IEEE Intelligent Systems*, 22(3):64–69.

[Huron, 2006] Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT Press.

[ISO/IEC, 2002] ISO/IEC (2002). 15938-5 information technology - multimedia content description interface - part 5: Multimedia description schemes. Technical report, FDIS.

[Jackendoff and Lerdahl, 2006] Jackendoff, R. and Lerdahl, F. (2006). The capacity for music: what is it, and what's special about it? *Cognition*, pages 33–72.

[Jaynes, 1957] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620–630.

[Jin et al., 2005] Jin, X., Mobasher, B., and Zhou, Y. (2005). A maximum entropy web recommendation system: combining collaborative and content features. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 612–617.

[Kelly, 1999] Kelly, K. (1999). *New Rules for the New Economy*. Penguin.

[Kolmogorov, 1965] Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems Information Transmission*, 1(1):1–7.

[Konstan et al., 1997] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J. (1997). Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87.

[Krumhansl, 2002] Krumhansl, C. (2002). Music: A link between cognition and emotion. *Current Directions in Psychological Science*, pages 35–55.

[Labov, 1973] Labov, W. (1973). *The Boundaries of Words and their Meanings*, pages 340–373. Oxford University Press, USA.

[Landauer and Dumais, 1997] Landauer, T. and Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, pages 211–240.

[Landauer et al., 1998a] Landauer, T. K., Foltz, P. W., and Laham, D. (1998a). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

[Landauer et al., 1998b] Landauer, T. K., Foltz, P. W., and Laham, D. (1998b). Latent semantic analysis passes the test: Knowledge representation and multiple-choise testing. Unpublished manuscript.

[Laurence and Margolis, 1999] Laurence, S. and Margolis, E. (1999). *Concepts and Cognitive Science*. MIT Press.

[Lekakos and Caravelas, 2006] Lekakos, G. and Caravelas, P. (2006). A hybrid approach for movie recommendation. *Multimedia Tools*, 36(55-70).

[Levitin and Menon, 2003] Levitin, D. J. and Menon, V. (2003). Musical structure is processed in "language" areas of the brain: a possible role for brodmann area 47 in temporal coherence. *Neuroimage*, 20(4):2142–2152.

[Levy and Sandler, 2007] Levy, M. and Sandler, M. (2007). A semantic space for music derived from social tags. *Proceedings of the 8th International Conference on Music Information Retrieval, Austrian Computer Society*, pages 411–416.

[Linden et al., 2003] Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations item-to-item collaborative filtering. *IEEE Internet Computing*, pages 76–80.

[Maddox, 1992] Maddox, W. T. (1992). Perceptual and decisional separability. *Multidimensional models of perception and cognition*, pages 147–180.

[Martinez et al., 2002] Martinez, J. M., Koenen, R., and Pereira, F. (2002). Mpeg-7: The generic multimedia content description standard, part 1. *IEEE*.

[McCloskey and Glucksberg, 1978] McCloskey, M. and Glucksberg, S. (1978). Natural categories: Well-defined or fuzzy sets. *Memory & Cognition*, 6(4):462–472.

[Medin and Schaffer, 1978] Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(1):207–238.

[Mervis et al., 1976] Mervis, C. B., Catlin, J., and Rosch, E. (1976). Relationships among goodness-of-example, category norms and word frequency. *Bulletin on Psychonomic Society*, 7:268–284.

[Meyer, 1957] Meyer, L. (1957). Meaning in music and information theory. *Journal of Aesthetics and Art Criticism*, pages 412–424.

[Miller et al., 2003] Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., and Riedl, J. (2003). Movielens unplugged: experiences with an occasionally connected recommender system. *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 263–266.

[Mobasher et al., 2007] Mobasher, B., Burke, R., Bhaumik, R., and Sandvig, J. (2007). Attacks and remedies in collaborative recommendation. *IEEE Intelligent Systems*, 22(3):56–63.

[Mobasher et al., 2000] Mobasher, B., Cooley, R., and Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of ACM*, 43(8):142–151.

[Nichols, 1997] Nichols, D. M. (1997). Implicit rating and filtering. *Proceedings of 5th DELOS Workshop on Filtering and Collaborative Filtering*, pages 31–36.

[NISO, 2004] NISO (2004). *Understanding Metadata*. National Information Standards Organization.

[Nosofsky, 1986] Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

[Nosofsky, 1987] Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1):87–108.

[Osgood et al., 1957] Osgood, C., Suci, G., and Tannenbaum, H. (1957). *The measurement of Meaning*. University of Illinois Press.

[Osinski and Weiss, 2005] Osinski, S. and Weiss, D. (2005). A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54.

[Pavlov et al., 2004] Pavlov, D., Manavoglu, E., Giles, C. L., and Pennock, D. M. (2004). Collaborative filtering with maximum entropy. *IEEE Intelligent Systems*, 19(6):40–48.

[Pazzani, 1997] Pazzani, M. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27(3):313–331.

[Pazzani, 1999] Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5):393–408.

[Pennock and Horvitz, 2000] Pennock, D. M. and Horvitz, E. (2000). Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 473–480.

[Peretz et al., 2004] Peretz, I., Radeau, M., and Arguin, M. (2004). Two-way interactions between music and language: Evidence from priming recognition of tune and lyrics in familiar songs. *Memory and Cognition*, pages 42–52.

[Petersen and Butkus, 2008a] Petersen, M. K. and Butkus, A. (2008a). Extracting moods from songs and bbc programs based on emotional context. *International Journal of Multimedia Broadcasting*, 2008(289837).

[Petersen and Butkus, 2008b] Petersen, M. K. and Butkus, A. (2008b). Modeling moods in bbc programs based on emotional context. In *LNCS Lecture Notes in Computer Science*.

[Petersen et al., 2008] Petersen, M. K., Hansen, L. K., Butkus, A., and Schwartz, M. (2008). Emotional vectors: Modelling media from cognitive components. *Submitted to the Journal of Multimedia Systems*.

[Pogacnik et al., 2005] Pogacnik, M., Tasic, J., Meza, M., and Kosir, A. (2005). Personal content recommender based on a hierarchical user model for the selection of tv programmes. *User Modeling and User-Adapted Interaction*, 15(5):425–457.

[Popesculand et al., 2001] Popesculand, A., Ungar, L. H., Pennock, D. M., and Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-2001)*.

[Prinz, 2002] Prinz, J. J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. MIT Press.

[Quah, 2003] Quah, D. (2003). Digital goods and the new economy. Discussion Paper No. 3846.

[Reisberg, 2001] Reisberg, D. (2001). *Cognition*. W. W. Norton and Company.

[Resnick et al., 1994] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). Grouplens: an open architecture for collaborative filtering of netnews. *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186.

[Resnick and Varian, 1997] Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of ACM*, 40(3):56–58.

[Rich, 1979] Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, 3:329–354.

[Richardson, 1964] Richardson, E. C. (1964). *Classification*. Hamden, CT: Shoe String Press, 3 edition.

[Rigg, 1937] Rigg, M. G. (1937). An experiment to determine how accuratelly college students can interpret intended meanings of musical compositions. *Journal of Experimental Psychology*, 21:223–229.

[Rips, 1975] Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, 14(6):665–681.

[Rips, 1989] Rips, L. J. (1989). Similarity, typicality, and categorization. *Similarity and analogical reasoning*, pages 21–59.

[Rosch, 1973] Rosch, E. (1973). *On the internal structure of perceptual and semantic categories*. New York: Academic Press.

[Rosch, 1975] Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192–233.

[Rosch, 1978] Rosch, E. (1978). Principles of categorization. *Cognition and categorization. Hillsdale, NJ: Erlbaum*.

[Rosch et al., 1976] Rosch, E., Mervis, C. B., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic objects in natural catergories. *Cognitive Psychology*, 3:382–439.

[Rucker and Polanco, 1997] Rucker, J. and Polanco, M. J. (1997). Personalized navigation for the web. *Communications of ACM*, 3(40):147–62.

[Salton, 1989] Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley.

[Sarwar et al., 2001] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th international conference on World Wide Web*, pages 285–295.

[Sarwar et al., 2000a] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. T. (2000a). Analysis of recommendation algorithms for e-commerce. *Proceedings of the 2nd ACM conference on Electronic commerce*.

[Sarwar et al., 2000b] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. T. (2000b). Application of dimensionality reduction in recommender system - a case study. *Proceedings of ACM WebKDD*.

[Schein et al., 2002] Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260.

[Schön et al., 2005] Schön, D., Gordon, R. L., and Besson, M. (2005). Musical and linguistic processing in song perception. *Annals of the New York Academy of Sciences*, pages 71–81.

[Sebe et al., 2003] Sebe, N., Lew, M. S., Zhou, X., Huang, T. S., and Bakker, E. M. (2003). The state of the art in image and video retrieval. *Image and Video Retrieval: Second International Conference*.

[Shahabi and Chen, 2003] Shahabi, C. and Chen, Y.-S. (2003). Web information personalization - challenges and approaches. *Databases in Networked Information Systems: Third International Workshop*.

[Shannon, 1948] Shannon, C. (1948). Mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

[Shapiro and Varian, 1999] Shapiro, C. and Varian, H. R. (1999). *Information Rules*. Harward business School Press.

[Shardanand and Maes, 1995] Shardanand, U. and Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth". *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 210–217.

[Shikhar, 1998] Shikhar, G. (1998). Making business sense of the internet. *Harvard Business Review*.

[Simon, 1971] Simon, H. A. (1971). Designing organizations for an information-rich world. *Computers, Communications and the Public Interest*, pages 38–52.

[Sinclair and Cardew-Hall, 2008] Sinclair, J. and Cardew-Hall, M. (2008). The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1).

[Smith, 1776] Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. University Of Chicago Press.

[Smith et al., 1974] Smith, E., Shoben, E. J., and Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81(3):214–241.

[Smith and Medin, 1981] Smith, E. E. and Medin, D. L. (1981). *Categories and concepts*. Harvard University Press Cambridge, Mass.

[Smyth and Cotter, 1999] Smyth, B. and Cotter, P. (1999). Surfing the digital wave. *Proceedings of the Third International Conference on Case-Based Reasoning and Development*, pages 561–571.

[Sollenborn and Funk, 2002] Sollenborn, M. and Funk, P. (2002). Category-based filtering and user stereotype cases to reduce the latency problem in recommender systems. *Advances in Case-Based Reasoning: 6th European Conference*, pages 395–420.

[Solomonoff, 1964] Solomonoff, R. (1964). A formal theory of inductive inference, part 1 and part 2. *Information and Control*, 7(2):224–254.

[Spiteri, 2007] Spiteri, L. F. (2007). The role of causality and conceptual coherence in assessments of similarity. *Library and Information Science Research Electronic Journal*, 17(2).

[Staab et al., 2002] Staab, S., Santini, S., Nack, F., Steels, L., and Maedche, A. (2002). Emergent semantics. *IEEE Intelligent Systems*, 17(1):78–86.

[Stam, 2000] Stam, R. (2000). *Film Theory: An Introduction*. Blackwell Publishers.

[Sullivan et al., 2004] Sullivan, D. O., Smyth, B., Wilson, D. C., McDonald, K., and Smeaton, A. (2004). Improving the quality of the personalized electronic program guide. *User Modeling and User-Adapted Interaction*, 14:5–36.

[Tapscott, 1999] Tapscott, D. (1999). *Creating Value in the Network Economy*. Harvard Business School Press.

[Tjondronegoro and Spink, 2008] Tjondronegoro, D. and Spink, A. (2008). Web search engine multimedia functionality. *Information Processing and Management*, 44:340–357.

[Tversky, 1977] Tversky, A. (1977). Features of similarity. *Psychological Review*, 84:327–352.

[Tzouvaras et al., 2007] Tzouvaras, V., Troncy, R., and Pan, J. Z. (2007). Multimedia annotation interoperability framework. Technical report, W3C Incubator Group Editor's Draft.

[Ungar and Foster, 1998] Ungar, L. H. and Foster, D. P. (1998). Clustering methods for collaborative filtering. *Workshop on recommender Systems at the 15th National Conference on Artificial Intelligence.*

[Varian, 2003] Varian, H. R. (2003). Economics of information technology. Technical report, University of California, Berkeley.

[Wactlar and Christel, 2002] Wactlar, H. D. and Christel, M. G. (2002). Digital video archives: Managing through metadata. *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving.*

[Wang et al., 2006] Wang, S., Mukherjee, S., and Anninger, S. (2006). Entertainment industry. the long tail. Technical report, Bear, Stearns and Co. Inc.

[WC3, 2004] WC3 (2004). Rdf primer. Technical report, WC3.

[Wikipedia, 2008] Wikipedia (2008). List of music styles.

[Wittgenstein, 1953] Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford: Blackwell.

[Yu and Zhou, 2004] Yu, Z. and Zhou, X. (2004). Tv3p - an adaptive assistant for personalized tv. *IEEE Transactions on Consumer Electronics*, 50(1):393–399.